

Digitized by the Internet Archive  
in 2023 with funding from  
University of Toronto

<https://archive.org/details/31761117123653>











12-001

138

---

# Survey Methodology

---

Catalogue No. 12-001-XPB

A journal  
published by  
Statistics Canada

June 2010

•

Volume 36

•

Number 1



Statistics  
Canada

Statistique  
Canada

Canada



## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at [www.statcan.gc.ca](http://www.statcan.gc.ca), e-mail us at [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca), or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

### Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

### Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

## To access and order this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website at [www.statcan.gc.ca](http://www.statcan.gc.ca) and select "Publications."

This product, Catalogue no. 12-001-X, is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca)
- Mail  
Statistics Canada  
Finance  
R.H. Coats Bldg., 6th Floor  
150 Tunney's Pasture Driveway  
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on [www.statcan.gc.ca](http://www.statcan.gc.ca) under "About us" > "Providing services to Canadians."



# Survey Methodology



A journal  
published by  
Statistics Canada

June 2010 • Volume 36 • Number 1

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2010

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows:  
Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

June 2010

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada  
Statistique Canada

Canada



# SURVEY METHODOLOGY

A Journal Published by Statistics Canada

*Survey Methodology* is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

## MANAGEMENT BOARD

**Chairman** J. Kovar

**Past Chairmen** D. Royce (2006-2009)  
G.J. Brackstone (1986-2005)  
R. Platek (1975-1986)

**Members** S. Fortier (Production Manager)  
J. Gambino  
M.A. Hidirolou  
J. Latimer  
H. Mantel

## EDITORIAL BOARD

**Editor** M.A. Hidirolou, *Statistics Canada*  
**Deputy Editor** H. Mantel, *Statistics Canada*

**Past Editor** J. Kovar (2006-2009)  
M.P. Singh (1975-2005)

## Associate Editors

J.M. Brick, *Westat Inc.*  
P. Cantwell, *U.S. Bureau of the Census*  
J.L. Eltinge, *U.S. Bureau of Labor Statistics*  
W.A. Fuller, *Iowa State University*  
J. Gambino, *Statistics Canada*  
D. Judkins, *Westat Inc.*  
D. Kasprzyk, *Mathematica Policy Research*  
P. Kott, *National Agricultural Statistics Service*  
P. Lahiri, *JPSM, University of Maryland*  
P. Lavallée, *Statistics Canada*  
G. Nathan, *Hebrew University*  
J. Opsomer, *Colorado State University*  
D. Pfeiffermann, *Hebrew University*  
N.G.N. Prasad, *University of Alberta*  
J.N.K. Rao, *Carleton University*

T.J. Rao, *Indian Statistical Institute*  
J. Reiter, *Duke University*  
L.-P. Rivest, *Université Laval*  
N. Schenker, *National Center for Health Statistics*  
F.J. Scheuren, *National Opinion Research Center*  
P. do N. Silva, *University of Southampton*  
E. Stasny, *Ohio State University*  
D. Steel, *University of Wollongong*  
L. Stokes, *Southern Methodist University*  
M. Thompson, *University of Waterloo*  
V.J. Verma, *Università degli Studi di Siena*  
K.M. Wolter, *Iowa State University*  
C. Wu, *University of Waterloo*  
A. Zaslavsky, *Harvard University*

**Assistant Editors** J.-F. Beaumont, C. Bocci, P. Dick, G. Dubreuil, S. Godbout, D. Haziza, Z. Patak, S. Rubin-Bleuer  
and W. Yung, *Statistics Canada*

## EDITORIAL POLICY

*Survey Methodology* publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

## Submission of Manuscripts

*Survey Methodology* is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

## Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

**Survey Methodology**  
A Journal Published by Statistics Canada  
Volume 36, Number 1, June 2010

**Contents**

Cochran-Hansen Prize 2011 .....	1
 <b>Regular Papers</b>	
Jiming Jiang, Thuan Nguyen and J. Sunil Rao Fence method for nonparametric small area estimation .....	3
Yan Lu and Sharon Lohr Gross flow estimation in dual frame surveys .....	13
Qixuan Chen, Michael R. Elliott and Roderick J.A. Little Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling .....	23
David Haziza, Katherine Jenny Thompson and Wesley Yung The effect of nonresponse adjustments on variance estimation .....	35
Jill A. Dever and Richard Valliant A comparison of variance estimators for poststratification to estimated control totals .....	45
Patrick J. Farrell and Sarjinder Singh Some contributions to jackknifing two-phase sampling estimators.....	57
Jason C. Legg and Cindy L. Yu A comparison of sample set restriction procedures .....	69
Mojca Bavdaž The multidimensional integral business survey response model .....	81
Lazarus Adua and Jeff S. Sharp Examining survey participation and response quality: The significance of topic salience and incentives .....	95
Tom Krenzke, Lin Li and Keith Rust Evaluating within household selection rules under a multi-stage design .....	111



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'«American National Standard for Information Sciences» – «Permanence of Paper for Printed Library Materials», ANSI Z39.48 - 1984.



## **Cochran-Hansen Prize 2011**

### **Competition for Young Survey Statisticians from Developing and Transitional Countries**

In celebration of its 25<sup>th</sup> anniversary, the International Association of Survey Statisticians (IASS) established the Cochran-Hansen Prize to be awarded every two years to the best paper on survey research methods submitted by a young statistician from a developing or transitional country.

Participation in the competition for the Prize is open to nationals of developing or transitional countries who are living in such countries and who were born in 1971 or later.

Papers submitted must be unpublished original works. They may include materials from the participant's university thesis. They should be in either English or French. Papers for consideration should be submitted to the IASS Secretariat at the address below to arrive by December 29, 2010. Each submission should be accompanied by a cover letter that gives the participant's year of birth, nationality, and country of residence. The cover letter must also indicate if the work submitted is the result of a PhD thesis and, in the case of joint papers, the prize candidate must state clearly what his/her contribution to the paper is.

The papers submitted will be examined by the Cochran-Hansen Prize Committee appointed by the IASS. The decision of the Committee is final.

The winner of the Prize will be invited to present his/her paper at the 58th Session of the International Statistical Institute to be held in Dublin, Ireland, August 21-29, 2011, and the name of the winner will be announced at the ISI General Assembly in Dublin.

The author of the winning paper will receive the Cochran-Hansen Prize in the form of books and journal subscriptions to the value of about € 500, and will have reasonable travel and living expenses paid in order to present the paper at the ISI Session in Dublin.

For further information, please contact:

Madame Claude OLIVIER  
IASS Secretariat  
International Association of Survey Statisticians  
CEFILINSEE, 3 rue de la Cité, 33500 Libourne, France  
Tel: +33 5 57 55 56 17  
Fax: +33 5 57 55 56 20  
E-mail: [Claude.olivier@insee.fr](mailto:Claude.olivier@insee.fr)







# Fence method for nonparametric small area estimation

Jiming Jiang, Thuan Nguyen and J. Sunil Rao<sup>1</sup>

## Abstract

This paper considers the problem of selecting nonparametric models for small area estimation, which recently have received much attention. We develop a procedure based on the idea of fence method (Jiang, Rao, Gu and Nguyen 2008) for selecting the mean function for the small areas from a class of approximating splines. Simulation results show impressive performance of the new procedure even when the number of small areas is fairly small. The method is applied to a hospital graft failure dataset for selecting a nonparametric Fay-Herriot type model.

Key Words: Fay-Herriot Model; Fence method; Nonparametric model selection; Penalized spline; Small area estimation.

## 1. Introduction

Small area estimation (SAE) has received increasing attention in recent literature. Here the term small area typically refers to a population for which reliable statistics of interest cannot be produced due to certain limitations of the available data. Examples of small areas include a geographical region (*e.g.*, a state, county, municipality, *etc.*), a demographic group (*e.g.*, a specific age  $\times$  sex  $\times$  race group), a demographic group within a geographic region, *etc.* In absence of adequate direct samples from the small areas, methods have been developed in order to “borrow strength”. Statistical models, especially mixed effects models, have played important roles in SAE. See Rao (2003) for a comprehensive account of various methods used in SAE.

While there is extensive literature on inference about small areas using mixed effects models, including estimation of small area means which is a problem of mixed model prediction, estimation of the mean squared error (MSE) of the empirical best linear unbiased predictor (EBLUP; see Rao 2003), and prediction intervals (*e.g.*, Chatterjee, Lahiri and Li 2007), model selection in SAE has received much less attention. However, the importance of model selection in SAE has been noted by prominent researchers in this field (*e.g.*, Battese, Harter and Fuller 1988, Ghosh and Rao 1994). Datta and Lahiri (2001) discussed a model selection method based on computation of the frequentist’s Bayes factor in choosing between a fixed effects model and a random effects model. They focused on the following one-way balanced random effects model for the sake of simplicity:  $y_{ij} = \mu + u_i + e_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, k$ , where the  $u_i$ ’s and  $e_{ij}$ ’s are normally distributed with mean zero and variances  $\sigma_u^2$  and  $\sigma_e^2$ , respectively. As noted by the authors, the choice between a

fixed effects model and a random effects one in this case is equivalent to testing the following one-sided hypothesis  $H_0: \sigma_u^2 = 0$  vs  $H_1: \sigma_u^2 > 0$ . Note that, however, not all model selection problems can be formulated as hypothesis testing. Fabrizi and Lahiri (2004) developed a robust model selection method in the context of complex surveys. Meza and Lahiri (2005) demonstrated the limitations of Mallows’  $C_p$  statistic in selecting the fixed covariates in a nested error regression model (Battese, Harter and Fuller 1988), defined as  $y_{ij} = x'_{ij}\beta + u_i + e_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ , where  $y_{ij}$  is the observation,  $x_{ij}$  is a vector of fixed covariates,  $\beta$  is a vector of unknown regression coefficients, and  $u_i$ ’s and  $e_{ij}$ ’s are the same as in the model above considered by Datta and Lahiri (2001). Simulation studies carried out by Meza and Lahiri (2005) showed that the  $C_p$  method without modification does not work well in the current mixed model setting when the variance  $\sigma_u^2$  is large; on the other hand, a modified  $C_p$  criterion developed by these latter authors by adjusting the intra-cluster correlations performs similarly as the  $C_p$  in regression settings. It should be pointed out that all these studies are limited to linear mixed models, while model selection in SAE in a generalized linear mixed model (GLMM) setting has never been seriously addressed.

Recently, Jiang *et al.* (2008) developed a new strategy for model selection, called *fence methods*. The authors noted a number of limitations of the traditional model selection strategies when applied to mixed model situations. For example, the BIC procedure (Schwarz 1978) relies on the effective sample size which is unclear in typical situations of SAE. To illustrate this, consider the nested error regression model introduced above. Clearly, the effective sample size is not the total number of observations  $n = \sum_{i=1}^m n_i$ , neither is proportional to  $m$ , the number of small areas unless all the  $n_i$  are equal and fixed. The fence methods avoid such

1. Jiming Jiang, University of California, Davis. E-mail: jiang@wald.ucdavis.edu; Thuan Nguyen, Oregon Health and Science University; J. Sunil Rao, Case Western Reserve University.



limitations, and therefore are suitable to mixed model selection problems, including linear mixed models and GLMMs. The basic idea of fence is to build a statistical fence to isolate a subgroup of what are known as the correct models. Once the fence is constructed, the optimal model is selected from those within the fence according to a criterion which can incorporate quantities of practical interest. More details about the fence methods are given below.

The focus of this paper is nonparametric models for SAE. These models have received much recent attention. In particular, Opsomer, Breidt, Claeskens, Kauermann and Ranalli (2007) proposed a spline-based nonparametric model for SAE. The idea is to approximate an unknown nonparametric small-area mean function by a penalized spline (P-spline). The authors then used a connection between P-splines and linear mixed models (Wand 2003) to formulate the approximating model as a linear mixed model, where the coefficients of the splines are treated as random effects. Consider, for simplicity, the case of univariate covariate. Then, a P-spline can be expressed as

$$\tilde{f}(x) = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + \dots + \gamma_q (x - \kappa_q)_+^p, \quad (1)$$

where  $p$  is the degree of the spline,  $q$  is the number of knots,  $\kappa_j, 1 \leq j \leq q$  are the knots, and  $x_+ = x1_{(x>0)}$ . Clearly, a P-spline is characterized by  $p, q$ , and also the location of the knots. Note that, however, given  $p, q$ , the location of the knots can be selected by the space-filling algorithm implemented in R [*cover.design()*]. But the question how to choose  $p$  and  $q$  remains. The general “rule of thumb” is that  $p$  is typically between 1 and 3, and  $q$  proportional to the sample size,  $n$ , with 4 or 5 observations per knot (Ruppert, Wand and Carroll 2003). But there may still be a lot of choices given the rule of thumb. For example, if  $n = 200$ , the possible choices for  $q$  range from 40 to 50, which, combined with the range of 1 to 3 for  $p$ , gives a total of 33 choices for the P-spline. Our new adaptive fence method offers a data-driven approach for choosing  $p$  and  $q$  for the spline-based SAE model.

The rest of the paper is organized as follows. The fence methods are described in section 2. In section 3 we develop an adaptive fence procedure for the nonparametric model selection problem. In section 4 we demonstrate the finite sample performance of the new procedure with a series of simulation studies. In section 5 we consider a real-life data example involving a dataset from a medical survey which has been used for fitting a Fay-Herriot model (Fay and Herriot 1979). Some technical results are deferred to the appendix.

## 2. Fence methods

As mentioned, the basic idea of fence is to construct a statistical fence and then select an optimal model from those within the fence according to certain criterion of optimality, such as model simplicity. Let  $Q_M = Q_M(y, \theta_M)$  be a measure of lack-of-fit, where  $y$  represents the vector of observations,  $M$  indicates a candidate model, and  $\theta_M$  denotes the vector of parameters under  $M$ . Here by lack-of-fit we mean that  $Q_M$  satisfies the basic requirement that  $E(Q_M)$  is minimized when  $M$  is a true model, and  $\theta_M$  the true parameter vector under  $M$ . Then, a candidate model  $M$  is in the fence if

$$\hat{Q}_M \leq \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M, \tilde{M}}, \quad (2)$$

where  $\hat{Q}_M = \inf_{\theta_M \in \Theta_M} Q_M$ ,  $\Theta_M$  being the parameter space under  $M$ ,  $\tilde{M}$  is a model that minimizes  $\hat{Q}_M$  among  $M \in \mathcal{M}$ , the set of candidate models, and  $\hat{\sigma}_{M, \tilde{M}}$  is an estimate of the standard deviation of  $\hat{Q}_M - \hat{Q}_{\tilde{M}}$ . The constant  $c_n$  on the right side of (2) can be chosen as a fixed number (e.g.,  $c_n = 1$ ) or adaptively (see below).

The calculation of  $\hat{Q}_M$  is usually straightforward. For example, in many cases  $Q_M$  can be chosen as the negative log-likelihood, or residual sum of squares. On the other hand, the computation of  $\hat{\sigma}_{M, \tilde{M}}$  can be quite challenging. Sometimes, even if an expression can be obtained for  $\hat{\sigma}_{M, \tilde{M}}$ , its accuracy as an estimate of the standard deviation cannot be guaranteed in a finite sample situation. Jiang, Nguyen and Rao (2009) simplified an adaptive fence procedure proposed by Jiang *et al.* (2008). For simplicity, we assume that  $\mathcal{M}$  contains a full model,  $M_f$ , of which each candidate model is a submodel. It follows that  $\tilde{M} = M_f$ . In the simplified adaptive procedure, the fence inequality (2) is replaced by

$$\hat{Q}_M - \hat{Q}_{M_f} \leq c_n, \quad (3)$$

where  $c_n$  is chosen adaptively as follows. For each  $M \in \mathcal{M}$ , let  $p^*(M) = P^*\{M_0(c) = M\}$  be the empirical probability of selection for  $M$ , where  $M_0(c)$  denotes the model selected by the fence procedure based on (3) with  $c_n = c$ , and  $P^*$  is obtained by bootstrapping under  $M_f$ . For example, under a parametric model one can estimate the model parameters under  $M_f$  and then use a parametric bootstrap to draw samples under  $M_f$ . Suppose that  $B$  samples are drawn, then  $p^*(M)$  is simply the sample proportion (out of a total of  $B$  samples) that  $M$  is selected by the fence procedure based on (3) with the given  $c_n$ . Let  $p^* = \max_{M \in \mathcal{M}} p^*(M)$ . Note that  $p^*$  depends on  $c_n$ . Let  $c_n^*$  be the  $c_n$  that maximizes  $p^*$  and this is our choice. Jiang *et al.* (2008) offers the following explanation of the motivation behind adaptive fence. Suppose that there is a true model among the candidate models, then, the optimal model is the one from which the data is generated, and

therefore should be the most likely given the data. Thus, given  $c_n$ , one is looking for the model (using the fence procedure) that is most supported by the data or, in other words, one that has the highest (posterior) probability. The latter is estimated by bootstrapping. Note that although the bootstrap samples are generated under  $M_f$ , they are almost the same as those generated under the optimal model. This is because the estimates corresponding to the zero parameters are expected to be close to zero, provided that the parameter estimators under  $M_f$  are consistent. One then pulls off the  $c_n$  that maximizes the (posterior) probability and this is the optimal choice.

There are two extreme cases corresponding to  $c_n = 0$  and  $c_n = \infty$  (i.e., very large). Note that if  $c_n = 0$ , then  $p^* = 1$ . This is because when  $c_n = 0$  the procedure always chooses  $M_f$ . Similarly, if there is a unique simplest model (e.g., model with minimum dimension), say,  $M_*$ , then  $p^* = 1$  for very large  $c_n$ . This is because, when  $c_n$  is large enough, all models are in the fence, hence the procedure always chooses  $M_*$ , if simplicity is used as the criterion of optimality for selecting the model within the fence. These two extreme cases are handled carefully in Jiang *et al.* (2008) and Jiang *et al.* (2009). However, as noted by Jiang *et al.* (2008), the procedures to handle the extreme cases, namely, the screen tests and baseline adjustment/threshold checking, are rarely needed in practice. For example, in most applications there are a (large) number of candidate variables, and it is believed that only a (small) subset of them are important. This means that the optimal model is neither  $M_*$  nor  $M_f$ . Therefore, there is no need to worry about the extreme cases, and the procedures to handle these cases can be skipped. In most applications a plot of  $p^*$  against  $c_n$  is W-shaped with the peak in the middle corresponding to  $c_n^*$ .

The left plot of Figure 2 provides an illustration. This is a plot of  $p^*$  against  $c_n$  for the example discussed in section 5. The plot shows the typical “W” shape, as described, and the peak in the middle corresponds to where the optimal  $c_n$ , i.e.,  $c_n^*$  is.

Jiang *et al.* (2009) established consistency of the simplified adaptive fence and studied its finite sample performance.

### 3. Nonparametric SAE model selection

For the simplicity of illustration we consider the following SAE model:

$$y_i = f(X_i) + B_i u_i + e_i, \quad i = 1, \dots, m, \quad (4)$$

where  $y_i$  is an  $n_i \times 1$  vector representing the observations from the  $i^{\text{th}}$  small area;  $f(X_i) = [f(x_{ij})]_{1 \leq j \leq n_i}$  with  $f(x)$  being an unknown (smooth) function;  $B_i$  is an  $n_i \times b$  known matrix;  $u_i$  is a  $b \times 1$  vector of small-area specific

random effects; and  $e_i$  is an  $n_i \times 1$  vector of sampling errors. It is assumed that  $u_i$ ,  $e_i$ ,  $i = 1, \dots, m$  are independent with  $u_i \sim N(0, G_i)$ ,  $G_i = G_i(\theta)$ , and  $e_i \sim N(0, R_i)$ ,  $R_i = R_i(\theta)$ ,  $\theta$  being an unknown vector of variance components. Note that, besides  $f(X_i)$ , the model is the same as the standard “longitudinal” linear mixed model (e.g., Laird and Ware 1982, Datta and Lahiri 2000).

The approximating spline model is given by replacing  $f(x)$  by  $\tilde{f}(x)$  in (1), where the coefficients  $\beta$ ’s and  $\gamma$ ’s are estimated by penalized least squares, i.e., by

$$\text{minimizing } |y - X\beta - Z\gamma|^2 + \lambda |\gamma|^2, \quad (5)$$

where  $y = (y_i)_{1 \leq i \leq m}$ , the  $(i, j)^{\text{th}}$  row of  $X$  is  $(1, x_{ij}, \dots, x_{ij}^p)$ , the  $(i, j)^{\text{th}}$  row of  $Z$  is  $[(x_{ij} - \kappa_1)_+^p, \dots, (x_{ij} - \kappa_q)_+^p]$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ , and  $\lambda$  is a penalty, or smoothing, parameter. To determine  $\lambda$ , Wand (2003) used the following interesting connection to a linear mixed model. To illustrate the idea, let us consider a simple case in which  $B_i = 0$  (i.e., there is no small-area random effects), and the components of  $e_i$  are independent and distributed as  $N(0, \tau^2)$ . If the  $\gamma$ ’s are treated as random effects which are independent and distributed as  $N(0, \sigma^2)$ , then the solution to (5) are the same as the best linear unbiased estimator (BLUE) for  $\beta$ , and the best linear unbiased predictor (BLUP) for  $\gamma$ , if  $\lambda$  is identical to the ratio  $\tau^2/\sigma^2$ . Thus, the value of  $\lambda$  may be estimated by the maximum likelihood (ML), or restricted maximum likelihood (REML) estimators of  $\sigma^2$  and  $\tau^2$  (e.g., Jiang 2007). However, there has been study suggesting that this approach is biased towards undersmoothing (Kauermann 2005). Consider, for example, a special case in which  $f(x)$  is, in fact, the quadratic spline with two knots given by (10). (Note that this function is smooth in that it has a continuous derivative.) It is clear that, in this case, the best approximating spline should be  $f(x)$  itself with only two knots, i.e.,  $q = 2$  (of course, one could use a spline with many knots to “approximate” the two-knot quadratic spline, but that would seem very inefficient in this case). However, if one uses the above linear mixed model connection, the ML (or REML) estimator of  $\sigma^2$  is consistent only if  $q \rightarrow \infty$  (i.e., the number of appearances of the spline random effects goes to infinity). The seeming inconsistency has two worrisome consequences: (i) the meaning of  $\lambda$  may be conceptually difficult to interpret; (ii) the behavior of the estimator of  $\lambda$  may be unpredictable.

The fence method offers a natural approach to choosing the degree of the spline,  $p$ , the number of knots,  $q$ , and the smoothing parameter,  $\lambda$  at the same time. Note, however, a major difference from the situations considered in Jiang *et al.* (2008) and Jiang *et al.* (2009) in that the true underlying model is not among the class of candidate models, i.e., the approximating splines (1). Furthermore, the



role of  $\lambda$  in the model should be made clear:  $\lambda$  controls the degree of smoothness of the underlying model. A natural measure of lack-of-fit is  $Q_M = |y - X\beta - Z\gamma|^2$ . However,  $\hat{Q}_M$  is not obtained by minimizing  $Q_M$  over  $\beta$  and  $\gamma$  without constraint. Instead, we have  $\hat{Q}_M = |y - X\hat{\beta} - Z\hat{\gamma}|^2$ , where  $\hat{\beta}$  and  $\hat{\gamma}$  are the solution to (5), and hence depends on  $\lambda$ . The optimal  $\lambda$  is to be selected by the fence method, together with  $p$  and  $q$ , as described below.

Another difference is that there may not be a full model among the candidate models. Therefore, the fence inequality (3) is replaced by the following:

$$\hat{Q}_M - \hat{Q}_{\tilde{M}} \leq c_n, \quad (6)$$

where  $\tilde{M}$  is the candidate model that has the minimum  $\hat{Q}_M$ . We use the following criterion of optimality within the fence which combines model simplicity and smoothness. For the models within the fence, choose the one with the smallest  $q$ ; if there are more than one such models, choose the model with the smallest  $p$ . This gives the best choice of  $p$  and  $q$ . Once  $p, q$  are chosen, we choose the model *within the fence* with the largest  $\lambda$ . Once again, note that  $\lambda$  is part of the model  $M$  that is selected (or “estimated”) by the fence method. The tuning constant  $c_n$  is chosen adaptively using the simplified adaptive procedure of Jiang *et al.* (2009), where parametric bootstrap is used for computing  $p^*$  (see section 2).

The following theorem is proved in Appendix. For simplicity, assume that the matrix  $W = (X \ Z)$  is of full rank. Let  $P_{W^\perp} = I_n - P_W$ , where  $n = \sum_{i=1}^m n_i$  and  $P_W = W(W'W)^{-1}W'$ .

**Theorem.** Computationally, the above fence procedure is equivalent to the following: (i) first use the (adaptive) fence to select  $p$  and  $q$  using (6) with  $\lambda = 0$  and  $\hat{Q}_M = y'P_{W^\perp}y$  (see Lemma below), and same criterion as above for choosing  $p, q$  within the fence; (ii) let  $M_0^*$  denotes the model corresponding to the selected  $p$  and  $q$ , find the maximum  $\lambda$  such that

$$\hat{Q}_{M_0^*, \lambda} - \hat{Q}_{\tilde{M}} \leq c_n^*, \quad (7)$$

where for any model  $M$  with the corresponding  $X$  and  $Z$ , we have

$$\hat{Q}_{M, \lambda} = |y - X\hat{\beta}_\lambda - Z\hat{\gamma}_\lambda|^2,$$

$$\hat{\beta}_\lambda = (X'V_\lambda^{-1}X)^{-1}X'V_\lambda^{-1}y,$$

$$\hat{\gamma}_\lambda = \lambda^{-1}(I_q + \lambda^{-1}Z'Z)^{-1}Z'(y - X\hat{\beta}_\lambda),$$

$$X'V_\lambda^{-1}X = X'X - \lambda^{-1}X'Z(I_q + \lambda^{-1}Z'Z)^{-1}Z'X,$$

$$X'V_\lambda^{-1}y = X'y - \lambda^{-1}X'Z(I_q + \lambda^{-1}Z'Z)^{-1}Z'y,$$

and  $c_n^*$  is chosen by the adaptive fence procedure described in section 2 ( $V_\lambda$  is defined below but not directly needed here for the computation because of the last two equations).

Note that in step (i) of the Theorem one does not need to deal with  $\lambda$ . The motivation for (7) is that this inequality is satisfied when  $\lambda = 0$ , so one would like to see how far  $\lambda$  can go. In fact, the maximum  $\lambda$  is a solution to the equation  $\hat{Q}_{M_0^*, \lambda} - \hat{Q}_{\tilde{M}} = c_n^*$ . The purpose of the last two equations is to avoid direct inversion of  $V_\lambda = I_n + \lambda^{-1}ZZ'$ , whose dimension is equal to  $n$ , the total sample size. Note that  $V_\lambda$  does not have a block diagonal structure because of  $ZZ'$ , so if  $n$  is large direct inversion of  $V_\lambda$  may be computationally burdensome.

The proof of the Theorem requires the following lemma, whose proof is given in Appendix.

**Lemma.** For any  $M$  and  $y$ ,  $\hat{Q}_{M, \lambda}$  is an increasing function of  $\lambda$  with  $\inf_{\lambda > 0} \hat{Q}_{M, \lambda} = \hat{Q}_M$ .

## 4. Simulations

We consider an extension of the Fay-Herriot model (Fay and Herriot 1979) in a nonparametric setting. The model can be expressed as

$$y_i = f(x_i) + v_i + e_i, \quad i = 1, \dots, m, \quad (8)$$

where  $v_i, e_i, i = 1, \dots, m$  are independent such that  $v_i \sim N(0, A)$ ,  $e_i \sim N(0, D_i)$ , where  $A$  is unknown but the sampling variance  $D_i$  is assumed known. The main difference from the traditional Fay-Herriot model is  $f(x_i)$ , where  $f(x)$  is an unknown smooth function.

For simplicity we assume  $D_i = D$ ,  $1 \leq i \leq m$ . Then, the model can be expressed as

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, m, \quad (9)$$

where  $\varepsilon_i \sim N(0, \sigma^2)$  with  $\sigma^2 = A + D$ , which is unknown. Thus, the model is the same as the nonparametric regression model.

We consider three different cases that cover various situations and aspects. In the first case, Case 1, the true underlying function is a linear function,  $f(x) = 1 - x$ ,  $0 \leq x \leq 1$ , hence the model reduces to the traditional Fay-Herriot model. The goal is to find out if fence can validate the traditional Fay-Herriot model in the case that it is valid. In the second case, Case 2, the true underlying function is a quadratic spline with two knots, given by

$$f(x) = 1 - x + x^2 - 2(x - 1)_+^2 + 2(x - 2)_+^2, \quad 0 \leq x \leq 3 \quad (10)$$

(the shape is half circle between 0 and 1 facing up, half circle between 1 and 2 facing down, and half circle between 2 and 3 facing up). Note that this function is smooth in that it has a continuous derivative. Here we intend to investigate whether the fence can identify the true underlying function in the “perfect” situation, *i.e.*, when  $f(x)$  itself is a spline. The last case, Case 3, is perhaps the most practical situation,

in which no spline can provide a perfect approximation to  $f(x)$ . In other words, the true underlying function is not among the candidates. In this case  $f(x)$  is chosen as  $0.5\sin(2\pi x)$ ,  $0 \leq x \leq 1$ , which is one of the functions considered by Kauermann (2005).

We consider situations of small or medium sample size, namely,  $m = 10, 15$  or  $20$  for Case 1,  $m = 30, 40$  or  $50$  for Case 2, and  $m = 10, 30$  or  $50$  for Case 3. The covariate  $x_i$  are generated from the Uniform $[0, 1]$  distribution in Case 1, and from Uniform $[0, 3]$  in Case 2; then fixed throughout the simulations. Following Kauermann (2005), we let  $x_i$  be the equidistant points in Case 3. The error standard deviation  $\sigma$  in (9) is chosen as  $0.2$  in Case 1 and Case 2. This value is chosen such that the signal standard deviation in each case is about the same as the error standard deviation. As for Case 3, we consider three different values for  $\sigma$ ,  $0.2, 0.5$  and  $1.0$ . These values are also of the same order as the signal standard deviation in this case.

The candidate approximating splines for Case 1 and Case 2 are the following:  $p = 0, 1, 2, 3$ ,  $q = 0$  and  $p = 1, 2, 3$ ,  $q = 2, 5$  (so there are a total of 10 candidates). As for Case 3, following Kauermann (2005), we consider only linear splines (*i.e.*,  $p = 1$ ); furthermore, we consider the number of knots in the range of the “rule of thumb” (*i.e.*, roughly 4 or 5 observations per knot; see section 1), plus the intercept model ( $p = q = 0$ ) and the linear model ( $p = 1, q = 0$ ). Thus, for  $m = 10$ ,  $q = 0, 2, 3$ ; for  $m = 30$ ,  $q = 0, 6, 7, 8$ ; and for  $m = 50$ ,  $q = 0, 10, 11, 12, 13$ .

Table 1 shows the results based on 100 simulations under Case 1 and Case 2. As in Jiang *et al.* (2009), we consider both

the highest peak, that is, choosing  $c_n$  with the highest  $p^*$ , and 95% lower bound (L.B.), that is, choosing a smaller  $c_n$  corresponding to a peak of  $p^*$  in order to be conservative, if the corresponding  $p^*$  is greater than the 95% lower bound of the  $p^*$  for any larger  $c_n$  that corresponds to a peak of  $p^*$ . It is seen that performance of the adaptive fence is satisfactory even with the small sample size. Also, it appears that the confidence lower bound method works better in smaller sample, but makes almost no difference in larger sample. These are consistent with the findings of Jiang *et al.* (2009).

**Table 1**

**Nonparametric model selection - Case 1 and Case 2. Reported are empirical probabilities, in terms of percentage, based on 100 simulations that the optimal model is selected**

Sample size	Case 1			Case 2		
	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$
Highest Peak	62	91	97	71	83	97
Confidence L.B.	73	90	97	73	80	96

Table 2 shows the results for Case 3. Note that, unlike Case 1 and Case 2, here there is no optimal model (an optimal model must be a true model according to our definition). So, instead of giving the empirical probabilities of selecting the optimal model, we give the empirical distribution of the selected models in each case. It is apparent that, as  $\sigma$  increases, the distribution of the models selected becomes more spread out. A reverse pattern is observed as  $m$  increases. The confidence lower bound method appears to perform better in picking up a model with splines. Within the models with splines, fence seems to overwhelmingly prefer fewer knots than more knots.

**Table 2**

**Nonparametric model selection - Case 3. Reported are empirical distributions, in terms of percentage, of the selected models**

Sample Size		$m = 10$		$m = 30$		$m = 50$	
# of Knots		$0, 2, 3$		$0, 6, 7, 8$		$0, 10, 11, 12, 13$	
		$(p, q)$	%	$(p, q)$	%	$(p, q)$	%
$\sigma = 0.2$	Highest Peak	(0, 0)	1	(1, 0)	9	(1, 10)	100
		(1, 0)	31	(1, 6)	91		
		(1, 2)	68				
	Confidence L.B.	(1, 0)	24	(1, 0)	9	(1, 10)	100
		(1, 2)	76	(1, 6)	91		
$\sigma = 0.5$	Highest Peak	(0, 0)	14	(1, 0)	21	(1, 0)	13
		(1, 0)	27	(1, 6)	77	(1, 10)	84
		(1, 2)	56	(1, 7)	2	(1, 11)	2
		(1, 3)	3			(1, 12)	1
	Confidence L.B.	(0, 0)	8	(1, 0)	8	(1, 0)	2
		(1, 0)	23	(1, 6)	89	(1, 10)	94
		(1, 2)	65	(1, 7)	3	(1, 11)	2
		(1, 3)	4			(1, 12)	2
		(0, 0)					
		(1, 0)					
		(1, 2)					
		(1, 3)					
$\sigma = 1$	Highest Peak	(0, 0)	27	(0, 0)	15	(0, 0)	10
		(1, 0)	20	(1, 0)	18	(1, 0)	26
		(1, 2)	49	(1, 6)	63	(1, 10)	60
		(1, 3)	4	(1, 7)	4	(1, 11)	2
	Confidence L.B.	(0, 0)	20	(0, 0)	1	(1, 12)	2
		(1, 0)	13	(1, 0)	13	(0, 0)	2
		(1, 2)	59	(1, 6)	82	(1, 0)	13
		(1, 3)	8	(1, 7)	4	(1, 10)	80
						(1, 11)	2
						(1, 12)	3



Note that the fence procedure allows us to choose not only  $p$  and  $q$  but also  $\lambda$  (see section 3). In each simulation we compute  $\hat{\beta} = \hat{\beta}_\lambda$  and  $\hat{\gamma} = \hat{\gamma}_\lambda$ , given below (7), based on the  $\lambda$  chosen by the adaptive fence. The fitted values are calculated by (1) with  $\beta$  and  $\gamma$  replaced by  $\hat{\beta}$  and  $\hat{\gamma}$ , respectively. We then average the fitted values over the 100 simulations. Figure 1 shows the average fitted values for the three cases ( $m=10, 30, 50$ ) with  $\sigma=0.2$  under Case 3. The true underlying function values,  $f(x_i) = 0.5 \sin(2\pi x_i)$ ,  $i=1, \dots, m$  are also plotted for comparison.

### 5. A real-life data example

We consider a dataset from Morris and Christiansen (1995) involving 23 hospitals (out of a total of 219 hospitals) that had at least 50 kidney transplants during a 27 month period (Table 3). The  $y_i$ 's are graft failure rates for kidney transplant operations, that is,  $y_i = \text{number of graft failures}/n_i$ , where  $n_i$  is the number of kidney transplants at hospital  $i$  during the period of interest. The variance for graft failure rate,  $D_i$ , is approximated by  $(0.2)(0.8)/n_i$ , where 0.2 is the observed failure rate for all hospitals. Thus,  $D_i$  is assumed known. In addition, a severity index  $x_i$  is available for each hospital, which is the average fraction of females, blacks, children and extremely ill kidney recipients at hospital  $i$ . The severity index is considered as a covariate.

**Table 3**  
**Hospital data from Morris and Christiansen (1995)**

Area	$y_i$	$x_i$	$\sqrt{D_i}$
1	0.302	0.112	0.055
2	0.140	0.206	0.053
3	0.203	0.104	0.052
4	0.333	0.168	0.052
5	0.347	0.337	0.047
6	0.216	0.169	0.046
7	0.156	0.211	0.046
8	0.143	0.195	0.046
9	0.220	0.221	0.044
10	0.205	0.077	0.044
11	0.209	0.195	0.042
12	0.266	0.185	0.041
13	0.240	0.202	0.041
14	0.262	0.108	0.036
15	0.144	0.204	0.036
16	0.116	0.072	0.035
17	0.201	0.142	0.033
18	0.212	0.136	0.032
19	0.189	0.172	0.031
20	0.212	0.202	0.029
21	0.166	0.087	0.029
22	0.173	0.177	0.027
23	0.165	0.072	0.025

Ganesh (2009) proposed a Fay-Herriot model for the graft failure rates, as follows:  $y_i = \beta_0 + \beta_1 x_i + v_i + e_i$ , where the  $v_i$ 's are hospital-specific random effects and  $e_i$ 's are sampling errors. It is assumed that  $v_i, e_i$  are independent with  $v_i \sim N(0, A)$  and  $e_i \sim N(0, D_i)$ . Here the variance

$A$  is unknown. Based on the model Ganesh obtained credible intervals for selected contrasts. However, inspections of the raw data suggest some nonlinear trends, which raises the question on whether the fixed effects part of the model can be made more flexible in its functional form.

To answer this question, we consider the Fay-Herriot model as a special member of a class of approximating spline models discussed in section 3. More specifically, we assume

$$y_i = f(x_i) + v_i + e_i, \quad i=1, \dots, m, \quad (11)$$

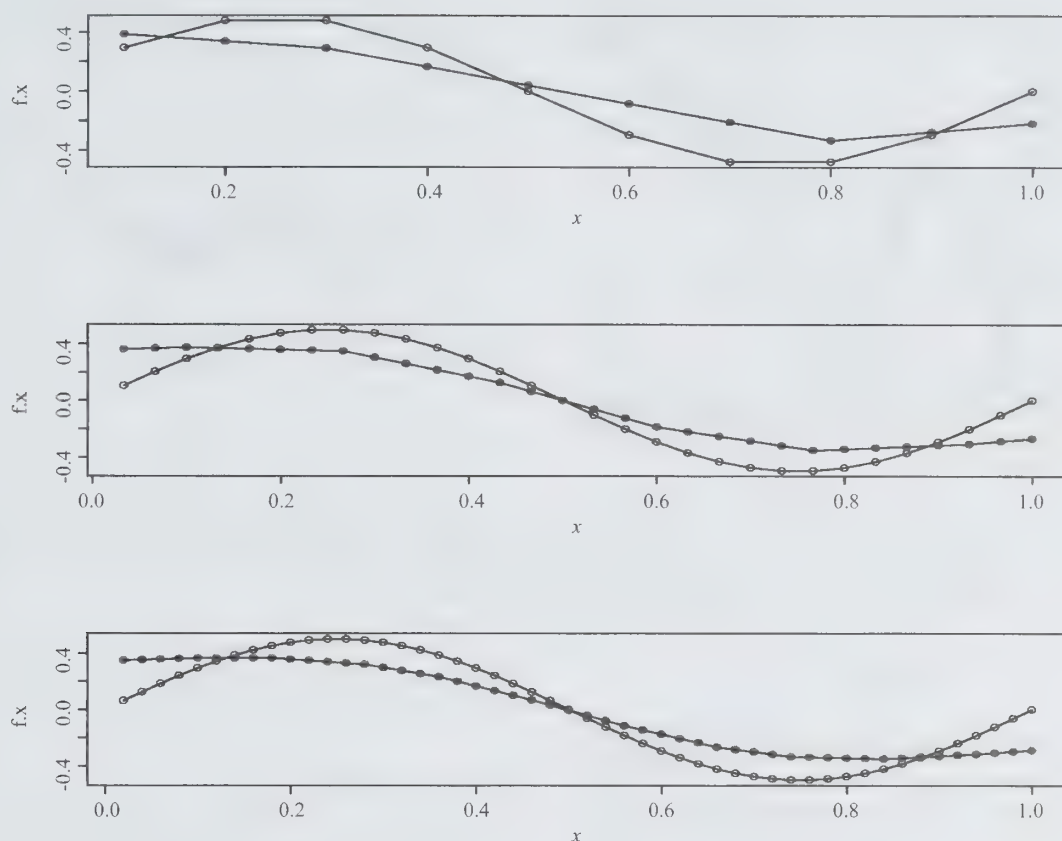
where  $f(x)$  is an unknown smooth function and everything else are the same as in the Fay-Herriot model. We then consider the following class of approximating spline models:

$$\hat{f}(x) = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + \dots + \gamma_q (x - \kappa_q)_+^p \quad (12)$$

with  $p=0, 1, 2, 3$  and  $q=0, 1, \dots, 6$  ( $p=0$  is only for  $q=0$ ). Here the upper bound 6 is chosen according to the "rule-of-thumb" (because  $m=23$ , so  $m/4=5.75$ ). Note that the Fay-Herriot model corresponds to the case  $p=1$  and  $q=0$ . The question is then to find the optimal model, in terms of  $p$  and  $q$ , from this class.

We apply the adaptive fence method described in section 3 to this case. Here to obtain the bootstrap samples needed for obtaining  $c_n^*$ , we first compute the ML estimator under the model  $\tilde{M}$ , which minimizes  $\hat{Q}_M = y' P_{W^\perp} y$  among the candidate models [i.e., (12); see Theorem in section 3], then draw parametric bootstrap samples under model  $\tilde{M}$  with the ML estimators treated as the true parameters. This is reasonable because  $\tilde{M}$  is the best approximating model in terms of the fit, even though under model (11) there may not be a true model among the candidate models. The bootstrap sample size is chosen as 100.

The fence method selects the model  $p=3$  and  $q=0$ , that is, a cubic function with no knots, as the optimal model. To make sure that the bootstrap sample size  $B=100$  is adequate, we repeated the analysis 100 times, each time using different bootstrap samples (recall in the adaptive fence one needs to draw bootstrap samples in order to determine  $c_n^*$ , so the question is whether different bootstrap samples lead to different results of model selection). All results led to the same model: a cubic function with no knots (even though the bootstrap-derived intermediate quantities, such as  $p^*$  and  $c_n^*$ , varied across bootstraps). We also ran the data analysis using  $B=1,000$ , and selected model remained the same. Thus, it appears that the bootstrap sample size  $B=100$  is adequate. The left figure of Figure 2 shows the plot of  $p^*$  against  $c_n$  in the adaptive fence model selection.



**Figure 1** Case 3 Simulation. Top figure: Average fitted values for  $m = 10$ . Middle figure: Average fitted values for  $m = 30$ . Bottom figure: Average fitted values for  $m = 50$ . In all cases, the dots represent the fitted values, while the circles correspond to the true underlying function

A few comparisons are always helpful. Our first comparison is to fence itself but with a more restricted space of candidate models. More specifically, we consider (12) with the restriction to linear splines only, *i.e.*,  $p=1$ , and knots in the range of the “rule of thumb”, *i.e.*,  $q=4, 5, 6$ , plus the intercept model ( $p=q=0$ ) and the linear model ( $p=1, q=0$ ). In this case, the fence method selected a linear spline with four knots (*i.e.*,  $p=1, q=4$ ) as the optimal model. The value of  $\lambda$  corresponding to this model is approximately equal to 0.001. The plot of  $p^*$  against  $c_n$  for this model selection is very similar to the left figure of Figure 2, and therefore omitted. In addition, the right figure of Figure 2 shows the fitted values and curves under the two models selected by the fence from within the different model spaces as well as the original data points.

A further comparison can be made by treating (11) as a generalized additive model (GAM) with heteroscedastic errors. A weighted fit can be obtained with the amount of smoothing optimized by using a generalized cross-validation (GCV) criterion. Here the weights used are  $w_i = 1/(A + D_i)$  where the maximum likelihood estimate for  $A$  is used as a plug-in estimate. Recall that the  $D_i$ 's are known. This fitted function is also overlayed in the right

figure of Figure 2. Notice how closely this fitted function resembles the restricted space fence fit.

To expand the class of models under consideration by GCV-based smoothing, we used the BRUTO procedure (Hastie and Tibshirani 1990) which augments the class of models to look at a null fit and a linear fit for the spline function; and embeds the resulting model selection (*i.e.*, null, linear or smooth fits) into a weighted backfitting algorithm using GCV for computational efficiency. Interestingly here, BRUTO finds simply an overall linear fit for the fixed effects functional form. While certainly an interesting comparison, BRUTO's theoretical properties for models like (11) have not really been studied in depth.

Finally, as mentioned in section 3, by using the connection between P-spline and linear mixed model one can formulate (12) as a linear mixed model, where the spline coefficients are treated as random effects. The problem then becomes a (parametric) mixed model selection problem, hence the method of Jiang *et al.* (2009) can be applied. In fact, this was our initial approach to this dataset, and the model we found was the same as the one by BRUTO. However, we have some reservation about this approach, as explained in section 3.



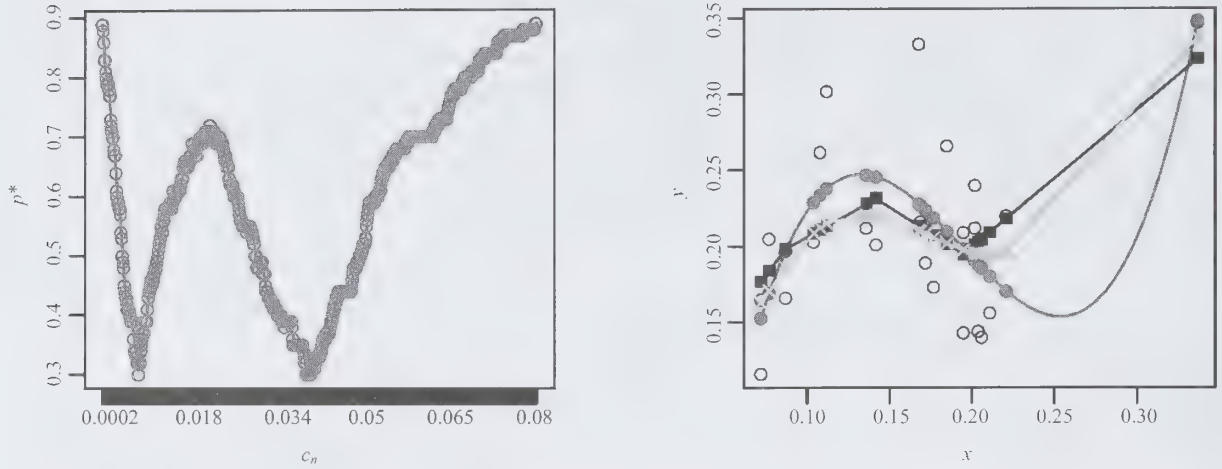


Figure 2 Left: A plot of  $p^*$  against  $c_n$  from the search over the full model space. Right: The raw data and the fitted values and curves; dots and their curve correspond to the cubic function resulted from the full model search; squares and their lines correspond to the linear spline with 4 knots resulted from the restricted model search; green 's and their lines represent the GAM fits

## 6. Concluding remarks

Although the focus of the current paper is nonparametric SAE model selection, our method may be applicable to spline-based mixed effects model selection problems in other areas, for example, in the analysis of longitudinal data (e.g., Wang 2005).

In the case where a true model exists among the candidate models, such as Cases 1 and 2 in section 4, consistency of the proposed fence model selection method can be established in the same way as in Section 3 of Jiang *et al.* (2009) (although the result of the latter paper does not directly apply). However, practically, the situation that nonparametric modeling is most useful is when a true model does not exist, or is not among the candidates, such as Case 3 in section 4. In this case, no result of consistency can be proved, of course. It remains unclear what is a desirable asymptotic behavior to study in the latter case.

## Acknowledgements

Jiming Jiang is partially supported by NSF grants DMS - 0203676 and DMS - 0402824. J. Sunil Rao is partially supported by NSF grants DMS - 0203724, DMS - 0405072 and NIH grant K25-CA89868.

## Appendix

1. *Proof of Lemma.* Write  $g(\lambda) = \hat{Q}_{M,\lambda}$ . It can be shown (detail omitted) that  $g'(\lambda) = 2\lambda y' B_\lambda A_\lambda B_\lambda' y$ , where  $A_\lambda = B'(W'W + \lambda BB')^{-1} B$ ,  $B_\lambda = W(W'W + \lambda BB')^{-1} B$  with  $B' = (0 \ I_q)$  and  $W = (X \ Z)$ . Hence  $g'(\lambda) \geq 0$  for  $\lambda > 0$ . Also  $\hat{Q}_{M,\lambda} \rightarrow \hat{Q}_M$  as  $\lambda \rightarrow 0$ .

2. *Proof of Theorem.* Consider the fence inequality

$$\hat{Q}_{M,\lambda} - \hat{Q}_{\bar{M},\bar{\lambda}} \leq c_n, \quad (\text{A.1})$$

where  $(\bar{M}, \bar{\lambda})$  minimizes  $\hat{Q}_{M,\lambda}$ . Also consider the fence inequality using  $\hat{Q}_M = y' P_{W^\perp} y$ , which is

$$\hat{Q}_M - \hat{Q}_{\bar{M}} \leq c_n. \quad (\text{A.2})$$

By Lemma, we must have  $\bar{\lambda} = 0$ , and  $\bar{M} = \tilde{M}$ , hence  $\hat{Q}_{\bar{M},\bar{\lambda}} = \hat{Q}_{\tilde{M}}$ . It follows, again by Lemma, that for the same  $c_n$ , (A.2) holds if and only if (A.1) holds for some  $\lambda$ . Therefore, the models within the fence, in terms of  $p$  and  $q$ , are the same under both procedures. It is then easy to see, according to the selection criterion, that the same model  $M_0 = M_0(c_n)$ , in terms of  $p$  and  $q$ , will be selected under both procedures for the given  $c_n$ . It then follows that the  $c_n^*$  selected using the adaptive procedure will be the same under both procedures. Then, once again using the above argument, the optimal model  $M_0^*$ , in terms of  $p$  and  $q$ , will be the same under both procedures.

The formulae below (7) can be derived using the expressions of BLUE and BLUP (e.g., Jiang 2007, §2.3.1) and the following identity (e.g., Sen and Srivastava 1990, page 275): If  $U$  is  $n \times q$  and  $V$  is  $q \times n$ , then  $(P + UV)^{-1} = P^{-1} - P^{-1}U(I_q + VP^{-1}U)^{-1}VP^{-1}$  so long as the inverses exist.

## References

- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 80, 28-36.

- Chatterjee, S., Lahiri, P. and Li, H. (2007). Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. *Annals of Statistics*, to appear.
- Datta, G.S., and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.
- Datta, G.S., and Lahiri, P. (2001). Discussions on a paper by Efron & Gous. (Ed., P. Lahiri) *Model Selection*, IMS Lecture Notes/Monograph 38.
- Fabrizi, E., and Lahiri, P. (2004). A new approximation to the Bayes information criterion in finite population sampling. Technical Report, Dept. of Math., Univ. of Maryland.
- Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Ganesh, N. (2009). Simultaneous credible intervals for small area estimation problems. *Journal of Multivariate Analysis*, in press.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.
- Hastie, T., and Tibshirani, R.J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.
- Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.
- Jiang, J., Rao, J.S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. *Annals of Statistics*, 36, 1669-1692.
- Jiang, J., Nguyen, T. and Rao, J.S. (2009). A simplified adaptive fence procedure. *Statistics and Probability Letters*, 79, 625-629.
- Kauermann, G. (2005). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planning and Inference*, 127, 53-69.
- Laird, N.M., and Ware, J.M. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963-974.
- Meza, J., and Lahiri, P. (2005). A note on the  $C_p$  statistic under the nested error regression model. *Survey Methodology*, 31, 105-109.
- Morris, C.N., and Christiansen, C.L. (1995). Hierarchical models for ranking and for identifying extremes with applications. *Bayes Statistics 5*, Oxford Univ. Press.
- Opsomer, J.D., Breidt, F.J., Claeskens, G., Kauermann, G. and Ranalli, M.G. (2007). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society B*, to appear.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Ruppert, R., Wand, M. and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Univ. Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Sen, A., and Srivastava, M. (1990). *Regression Analysis*. New York: Springer.
- Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18, 223-249.
- Wang, J.-L. (2005). Nonparametric regression analysis of longitudinal data. *Encyclopedia of Biostatistics*, 2<sup>nd</sup> Ed.





# Gross flow estimation in dual frame surveys

Yan Lu and Sharon Lohr<sup>1</sup>

## Abstract

Gross flows are often used to study transitions in employment status or other categorical variables among individuals in a population. Dual frame longitudinal surveys, in which independent samples are selected from two frames to decrease survey costs or improve coverage, can present challenges for efficient and consistent estimation of gross flows because of complex designs and missing data in either or both samples. We propose estimators of gross flows in dual frame surveys and examine their asymptotic properties. We then estimate transitions in employment status using data from the Current Population Survey and the Survey of Income and Program Participation.

**Key Words:** Complex surveys; Dual frame surveys; Jackknife; Longitudinal estimation; Missing data.

## 1. Introduction

Many current surveys follow the same individuals at regular time intervals so that longitudinal quantities such as transitions in employment status and poverty status can be studied. The U.S. Current Population Survey (CPS; United States Census Bureau 2006), for example, uses a rotating panel design in which persons in a housing unit selected for the survey are interviewed for four consecutive months, rested for eight months, and then interviewed again for four consecutive months. This design allows estimation of quantities related to individuals' changes over time. Since many survey responses are categorical, gross flows, which are transitions among states of a categorical variable over time, are particularly important.

Table 1 displays the counts of a categorical variable measured at two times in a population of  $N$  units. At time 1, the variable can be in one of  $r$  states and at time 2, the variable can be in one of  $c$  states. To illustrate Table 1, we give the following example. In studying changes in employment status, we might have  $r = 2$  and  $c = 2$ , with state 0 representing unemployment and state 1 representing employment. Then  $X_{00}$  gives the count of persons in the population who are unemployed at both times,  $X_{10}$  is the number of persons who are employed at time 1 but unemployed at time 2,  $X_{0+}$  is the total number of persons who are unemployed at time 1, and so on. It is of interest to obtain estimates and standard errors of the gross flows  $X_{kl}$ ,  $k = 0, \dots, r-1$ ,  $l = 0, \dots, c-1$ , using survey data. This can be complicated in practice because of missing data and other problems.

While successive cross-sectional estimates can assess a change in unemployment rates over time, only a longitudinal survey addresses issues such as persistence of unemployment in individuals. Gross flow estimation using survey data has been studied by many authors, including

Chambers, Woyzbun and Pillig (1988), Hocking and Oxspring (1971), Blumenthal (1968), Chen and Fienberg (1974), Stasny (1984, 1987), and Stasny and Fienberg (1986). Most of this work considered methods for obtaining maximum likelihood (ML) estimators for expected cell values in contingency tables with partially cross-classified data. Pfeffermann, Skinner and Humphreys (1998) proposed estimators that account for misclassification in survey data. All of this work has assumed that a probability sample, usually a simple random sample, has been taken from a single sampling frame.

**Table 1**  
**Gross flow table for population**

		Time 2					
		0	1	2	...	$c-1$	
Time 1	0	$X_{00}$	$X_{01}$	$X_{02}$	...	$X_{0,c-1}$	$X_{0+}$
	1	$X_{10}$	$X_{11}$	$X_{12}$	...	$X_{1,c-1}$	$X_{1+}$
	2	$X_{20}$	$X_{21}$	$X_{22}$	...	$X_{2,c-1}$	$X_{2+}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$r-1$	$X_{r-1,0}$	$X_{r-1,1}$	$X_{r-1,2}$	...	$X_{r-1,c-1}$	$X_{r-1,+}$
		$X_{+0}$	$X_{+1}$	$X_{+2}$	...	$X_{+,c-1}$	$N$

A number of longitudinal surveys, such as the Canadian National Longitudinal Survey of Children and Youth and the Canadian Household Panel Survey, have now started or are considering implementation of a dual frame or multiple frame design. In a multiple frame survey, probability samples are selected independently from two or more frames. Using more than one frame often gives better coverage of the population, and can achieve considerable cost savings in some populations. For example, the Assets and Health Dynamics Survey (Heeringa 1995), with the goal of estimating characteristics of the population aged over 65, used a dual frame survey in which frame  $A$  was

1. Yan Lu, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001. E-mail: yljenlu@gmail.com;  
Sharon Lohr, School of Mathematical and Statistical Sciences, Arizona State University, Tempe AZ 85287-1804. E-mail: sharon.lohr@asu.edu.



the frame for a national general population survey and frame  $B$  was a list of Medicare enrollees. The structure of this survey is illustrated in Figure 1. Frame  $A$  covered the entire population but required extensive screening to identify individuals in the target population and was thus expensive to sample from; frame  $B$  was less expensive to sample, but did not include the entire population. Kalton and Anderson (1986) described uses of dual frame surveys to sample rare populations; Blair and Blair (2006) argued that dual frame surveys can take advantage of less expensive sampling modes such as internet sampling when sampling rare populations.

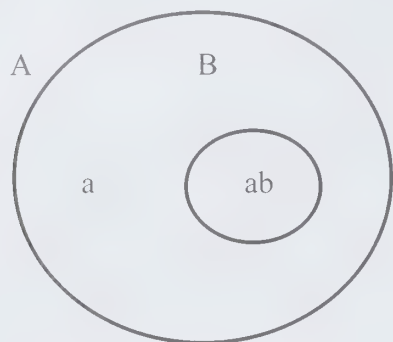


Figure 1 Frame  $B$  is a subset of frame  $A$

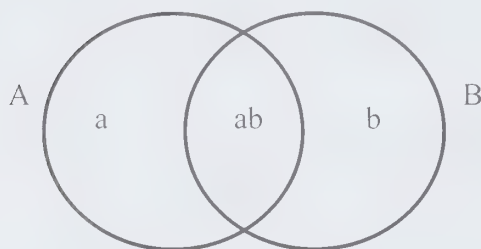


Figure 2 Frames  $A$  and  $B$  are both incomplete but overlapping

In other situations, both frames may be incomplete, as depicted in Figure 2. Hartley (1962, 1974) first proposed estimators for the dual frame survey design in Figure 2, when independent samples are taken from each frame. Subsequent developments are given in Bankier (1986), Fuller and Burmeister (1972), Skinner and Rao (1996), and Lohr and Rao (2000). Lohr and Rao (2006) summarized methods for estimating population quantities in cross-sectional multiple frame surveys.

In this paper, we propose estimators for gross flows that can be applied to dual frame surveys in which longitudinal information is collected in one or both samples. Units sampled in one or both surveys are followed over time; in some cases, additional units are sampled at later times to incorporate new population units or compensate for attrition. A longitudinal dual frame survey presents additional challenges to those found in longitudinal single frame

surveys or in cross-sectional dual frame surveys. Missing data can occur in the sample from either frame, and units may change frame membership between interviews in the survey. In addition, either sampling design may be complex, with stratification and clustering. In an overlapping dual frame survey such as that depicted in Figure 2, one wishes to use the information in the overlap as efficiently as possible. The problem studied in this article is to use all the information sampled from frame  $A$  and frame  $B$  to estimate the transition probabilities of the population.

The article is organized as follows. In Section 2, we set up the research problem. In Section 3, we derive gross flow estimators in dual frame surveys for complex samples with possibly missing data. In Section 4, we derive asymptotic properties and discuss variance estimation. An application of our research to the Current Population Survey and Survey of Income and Program Participation is given in Section 5. Finally, we give our conclusions in Section 6.

## 2. Notation and sample quantities

Suppose there are two sampling frames, frame  $A$  and frame  $B$ , which together cover the population of interest  $A \cup B$  as shown in Figure 2. In Hartley's (1962) notation, there are three nonoverlapping domains:  $a = A \cap B^c$ ,  $b = A^c \cap B$ , and  $ab = A \cap B$ , where  $c$  denotes complement of a set. The population sizes for frames  $A$  and  $B$  are  $N_A$  and  $N_B$ , with domain population sizes  $N_a$ ,  $N_b$ , and  $N_{ab}$ . We assume that  $N_A$  and  $N_B$  are known, but the population size  $N = N_A + N_B - N_{ab}$  may be unknown. In this article, we assume that both the population and the frames are fixed over time. These are strong assumptions but in many longitudinal surveys the population of interest and the frames may be defined for time 1.

Assume for this section that domain membership is constant over time. For simplicity of notation in this paper we assume that  $r = 2$  and  $c = 2$  so that there are two possible categories at each time; the general case is similar. Since the three domains are nonoverlapping, each population count  $X_{kl}$ ,  $k = 0, 1$ ,  $l = 0, 1$ , can be written as  $X_{kl} = X_{kla} + X_{klab} + X_{klb}$ , where  $X_{kld}$  is the number of population units in domain  $d$  that are in state  $k$  at time 1 and state  $l$  at time 2. The corresponding population and domain probabilities are  $p_{kl} = X_{kl}/N$  and  $p_{kld} = X_{kld}/N_d$  for  $d \in \{a, ab, b\}$ .

Independent probability samples,  $S_A$  and  $S_B$ , with sample sizes  $n_A$  and  $n_B$ , are taken from frames  $A$  and  $B$ . Let  $w_i^A$  be the weight of sampled unit  $i$  for the sample from frame  $A$  and let  $w_j^B$  be the weight of sampled unit  $j$  for the sample from frame  $B$ . We may take  $w_i^A$  to be the sampling weight  $[P(i \in S_A)]^{-1}$  or a Hájek-type weight  $[P(i \in S_A)]^{-1} N_A / (\text{sum of sampling weights in } S_A)$ . Other

weighting schemes for longitudinal data, discussed in Verma, Betti and Ghellini (2007) and Lavallée (2007), might also be used. Let  $\mathbf{y}_i = (y_{i1}, y_{i2})$  be the response for unit  $i$  in  $S_A$ , with  $y_{i1}, y_{i2} \in \{0, 1, M\}$  where  $M$  denotes that the value is missing. Then  $\hat{X}_{kla}^A = \sum_{i \in S_A} w_i^A I(y_{i1} = k) I(y_{i2} = l) I(i \in a)$  and  $\hat{X}_{klab}^A = \sum_{i \in S_A} w_i^A I(y_{i1} = k) I(y_{i2} = l) I(i \in ab)$  estimate the population counts for the  $(k, l)$  cell in domains  $a$  and  $ab$  from  $S_A$ , for  $k, l \in \{0, 1, M\}$ . Let  $\mathbf{y}_j = (y_{j1}, y_{j2})$  be the response for unit  $j$  in  $S_B$ , and let  $\hat{X}_{klb}^B = \sum_{j \in S_B} w_j^B I(y_{j1} = k) I(y_{j2} = l) I(j \in b)$  and  $\hat{X}_{klab}^B = \sum_{j \in S_B} w_j^B I(y_{j1} = k) I(y_{j2} = l) I(j \in ab)$  be the corresponding estimators from  $S_B$ .

In this paper, we assume that domain membership can be determined for every sample unit and that the responses  $\mathbf{y}_i$  have no classification error. Thus, we assume that we know whether each unit in the frame  $A$  or frame  $B$  sample belongs to the other frame or not. We also assume that there is no measurement error for  $\mathbf{y}_i$  and  $\mathbf{y}_j$  – in the employment example, this means that every respondent gives the correct response for his or her employment status. Thus, the methods we proposed in our article are sensitive to misclassification of observations into domains and into cells. If the domain means differ or if observations are classified incorrectly, the estimators of gross flows could be biased; Pfeffermann *et al.* (1998) discussed methods of accounting for misclassification in single frame surveys.

The estimators from  $S_A$  are displayed in Table 2. A similar table may be constructed for the estimators from  $S_B$ . We assume that each unit is sampled during one or both time periods. If there is no missing data, then all the estimated counts for cells  $(k, M)$  and  $(M, l)$  are zero. Using the exact or approximate unbiasedness of the estimators, depending on whether the sampling or Hájek weights are used, when there is no missing data,  $E[\hat{X}_{kla}^A] \approx X_{kla}$ ,  $E[\hat{X}_{klab}^A] \approx E[\hat{X}_{klab}^B] \approx X_{klab}$  and  $E[\hat{X}_{klb}^B] \approx X_{klb}$ .

**Table 2**  
**Estimators from the frame  $A$  sample**

		Time 2				
		0	1	Missing		
Time 1	domain $a$	0	$\hat{X}_{00a}^A$	$\hat{X}_{01a}^A$	$\hat{X}_{0Ma}^A$	$\hat{X}_{0+a}^A$
		1	$\hat{X}_{10a}^A$	$\hat{X}_{11a}^A$	$\hat{X}_{1Ma}^A$	$\hat{X}_{1+a}^A$
		Missing	$\hat{X}_{M0a}^A$	$\hat{X}_{M1a}^A$		$\hat{X}_{M+a}^A$
	domain $ab$	0	$\hat{X}_{00ab}^A$	$\hat{X}_{01ab}^A$	$\hat{X}_{0Mab}^A$	$\hat{X}_{0+ab}^A$
		1	$\hat{X}_{10ab}^A$	$\hat{X}_{11ab}^A$	$\hat{X}_{1Mab}^A$	$\hat{X}_{1+ab}^A$
		Missing	$\hat{X}_{M0ab}^A$	$\hat{X}_{M1ab}^A$		$\hat{X}_{M+ab}^A$
		$\hat{X}_{+0}^A$	$\hat{X}_{+1}^A$	$\hat{X}_{+M}^A$	$\hat{N}_A$	

### 3. Gross flow estimators in dual frame surveys

In this section, we derive gross flow estimators for complex samples in dual frame surveys. A dual frame pseudo-likelihood approach is used to account for the sampling designs and missing data mechanism. A dual frame approach can improve precision of the estimators and provide more flexibility to model the missing data mechanism. Methods in current use for handling missing data are based on standard statistical methods and fall into four general categories (Little and Rubin 2002): complete-case analysis, weighting methods, imputation methods and model-based methods. We adopt a model-based approach for the missing data. In this section, we first consider a simple setup with simple random samples from a population with no missing data. Then we add a model for the missing data mechanism. Finally, we discuss estimators for more complex survey designs.

#### 3.1 Simple random samples with complete data

To motivate the estimator in the general case, we first study estimation of gross flows when there is no missing data and when the sample from each frame is a simple random sample. Then  $x_{kld}^A = n_A \hat{X}_{kld}^A / N_A$ , for  $d = a, ab$ , is the observed sample count in cell  $kl$  and domain  $d$  from  $S_A$ ;  $x_{kld}^B = n_B \hat{X}_{kld}^B / N_B$  for  $d = b, ab$  is the corresponding observed sample count from  $S_B$ .

If the sampling fractions are small, a multinomial approximation may be used for the likelihood. For the sample from frame  $A$ , there are eight cells with associated probabilities  $P_{kld}^A = p_{kld} N_d / N_A$ , for  $k, l \in \{0, 1\}$  and  $d \in \{a, ab\}$ . The related probabilities for the sample from frame  $B$  are  $P_{kld}^B = p_{kld} N_d / N_B$  for  $k, l \in \{0, 1\}$  and  $d \in \{b, ab\}$ . Using the multinomial distribution and the assumption that the samples from the two frames are selected independently, the likelihood function is

$$L(\mathbf{p}, N_{ab}) \propto \prod_{k,l,d} (P_{kld}^A)^{x_{kld}^A} \times \prod_{k,l,d} (P_{kld}^B)^{x_{kld}^B}.$$

Although the likelihood is written for simplicity in terms of  $P_{kld}^A$  and  $P_{kld}^B$ , the underlying parameters of interest are  $\mathbf{p} = (p_{00a}, p_{01a}, \dots, p_{11b})$  and  $N_{ab}$ .

Setting the partial derivatives of the loglikelihood with respect to the parameters equal to zero, the maximum likelihood estimators are  $\hat{p}_{kla} = x_{kla} / n_a$ ,  $\hat{p}_{klb} = x_{klb} / n_b$  and  $\hat{p}_{klab} = (x_{kla}^A + x_{klab}^B) / (n_{ab}^A + n_{ab}^B)$ , where  $n_{ab}^A = \sum_{i \in S_A} I(i \in ab)$ ,  $n_{ab}^B = \sum_{j \in S_B} I(j \in ab)$ ,  $n_a^A = n_A - n_{ab}^A$  and  $n_b^B = n_B - n_{ab}^B$ . The MLE for  $N_{ab}$ ,  $\hat{N}_{ab}$ , is the smaller root of the quadratic equation

$$[n_a^A + n_b^B] \hat{N}_{ab}^2 - [n_a^A N_B + n_b^B N_A + n_{ab}^A N_A + n_{ab}^B N_B] \hat{N}_{ab} + [n_{ab}^A + n_{ab}^B] N_A N_B = 0. \quad (1)$$



Finally, using the above results, we construct the MLEs for  $X_{kl}$  and  $p_{kl}$ :

$$\begin{aligned}\hat{X}_{kl} &= (N_A - \hat{N}_{ab}) \hat{p}_{kla} + \hat{N}_{ab} \hat{p}_{klab} + (N_B - \hat{N}_{ab}) \hat{p}_{klb} \\ \hat{p}_{kl} &= \frac{(N_A - \hat{N}_{ab}) \hat{p}_{kla} + \hat{N}_{ab} \hat{p}_{klab} + (N_B - \hat{N}_{ab}) \hat{p}_{klb}}{N_l + N_B - \hat{N}_{ab}}.\end{aligned}$$

These estimators are the same as those obtained by Skinner (1991). However, Skinner used the approximate normal distribution of the response mean  $\bar{y}$  in each domain to obtain the MLEs, while our estimators come from a multinomial model. The multinomial model allows us to include partially classified information from units observed at only one time period, as shown in the next section.

### 3.2 Simple random samples with missing data

In practice, individuals may appear in the sample at only one of the times. This can occur due to sample attrition (when members of the sample drop out during the course of a study) or other causes. In a rotating panel survey such as the CPS, persons rotating out of the survey at time 1 will not be contacted for time 2 and thus their time-2 employment status will be unknown. In other situations, one of the samples may be cross sectional, in which case all observations are measured at exactly one time.

#### 3.2.1 Model for missing data

Blumenthal (1968), Chen and Fienberg (1974), Stasny (1984, 1987) and Stasny and Fienberg (1986) used a two-phase procedure to model the missing data in a single sample. A model is proposed for the complete data, and then the missing data mechanism is modeled. We extend this procedure to our dual frame structures. One advantage of a dual frame survey is that it provides more flexibility for the missing data models.

First, we assume that if all units were measured at both times, the model in Section 3.1 could be used. For the non-response mechanism, assume that each observation in cell  $(k, l)$  and domain  $d$  from  $S_A$  has probability  $\phi_{kld}^A$  of being missing at time 1 and probability  $\psi_{kld}^A$  of being missing at time 2. We assume the unit cannot be missing at both times.

This formulation assumes a constant probability that an observation will be missing within a given cell, domain, and frame. If data could be missing for different reasons, additional parameters could be used to distinguish observations that have partial classification because of, say, the rotating panel design, and observations that have partial classification because of nonresponse. In Section 5, we discuss an alternative approach that might be used with multiple mechanisms for missing data.

For  $k, l \in \{0, 1\}$ , the probability that a unit from  $S_A$  is observed in cell  $(k, l)$  and domain  $d$  is

$$Q_{kld}^A = P_{kld}^A (1 - \phi_{kld}^A - \psi_{kld}^A).$$

The probability that a unit from  $S_A$  is observed in cell  $(k, M)$  and domain  $d$  is

$$Q_{kMd}^A = \sum_{l=0}^1 P_{kld}^A \psi_{kld}^A.$$

Similarly, the probability that a unit from  $S_A$  is observed in cell  $(M, l)$  and domain  $d$  is

$$Q_{Mld}^A = \sum_{k=0}^1 P_{kld}^A \phi_{kld}^A.$$

The probabilities for frame  $B$  are defined similarly with  $Q_{kld}^B = P_{kld}^B (1 - \phi_{kld}^B - \psi_{kld}^B)$ ,  $Q_{kMd}^B = \sum_{l=0}^1 P_{kld}^B \psi_{kld}^B$  and  $Q_{Mld}^B = \sum_{k=0}^1 P_{kld}^B \phi_{kld}^B$ .

Under this two phase model, and using the assumption of independence of the samples, the likelihood function for the two samples is:

$$\begin{aligned}L(\mathbf{p}, \boldsymbol{\psi}, \boldsymbol{\phi}, N_{ab}) &\propto \prod_{k \in \{0, 1\}} \prod_{l \in \{0, 1\}} \prod_{d \in \{a, ab\}} (Q_{kld}^A)^{x_{kld}^A} \\ &\times \prod_{k \in \{0, 1\}} \prod_{l \in \{0, 1\}} \prod_{d \in \{b, ab\}} (Q_{kld}^B)^{x_{kld}^B} \\ &\times \prod_{k \in \{0, 1\}} \prod_{d \in \{a, ab\}} (Q_{kMd}^A)^{x_{kMd}^A} \\ &\times \prod_{l \in \{0, 1\}} \prod_{d \in \{a, ab\}} (Q_{Mld}^A)^{x_{Mld}^A} \\ &\times \prod_{k \in \{0, 1\}} \prod_{d \in \{b, ab\}} (Q_{kMd}^B)^{x_{kMd}^B} \\ &\times \prod_{l \in \{0, 1\}} \prod_{d \in \{b, ab\}} (Q_{Mld}^B)^{x_{Mld}^B},\end{aligned}\quad (2)$$

where  $\boldsymbol{\psi}$  is the vector of  $\psi_{kld}^A$ 's and  $\psi_{kld}^B$ 's and  $\boldsymbol{\phi}$  is the vector of  $\phi_{kld}^A$ 's and  $\phi_{kld}^B$ 's.

The expression in (2) is for the most general model, in which both surveys are longitudinal and both have missing data at each time period. If frame  $A$  uses a rotating panel survey, for example, then all of the probabilities  $Q_{kld}^A$  are nonzero: the units in the panels measured at both time periods will be included in the estimators  $x_{kld}^A$  for  $k, l \in \{0, 1\}$ , the units in the panels leaving the survey after time 1 will be included in the estimators  $x_{kMd}^A$ , and the units in the incoming panels will be included in the estimators  $x_{Mld}^A$ . Depending on the structure of the surveys, some of the factors in (2) may be omitted. For example, if the survey from frame  $B$  is a repeated cross-sectional survey with small sampling fraction, the probabilities  $Q_{kld}^B$  for  $k, l \in \{0, 1\}$  will be close to zero, and we would omit those factors from the likelihood.

The likelihood in (2) can be written as a product of a factor with  $N_{ab}$  and a factor containing the remaining parameters. As a consequence, the MLE for  $N_{ab}$  is again the smaller root of the equation in (1). We discuss the estimators of the remaining parameters in the next section.

### 3.2.2 Model identifiability and reduced models

A problem with maximizing the likelihood in (2) is that under the general model there are a total of 42 parameters while the two samples have only 32 observed cell counts. Thus we cannot estimate all the parameters under the most general model. But we can consider models with reduced parameterizations, as done in Chen and Fienberg (1974) for single frame surveys. The dual frame situation, in fact, gives much more flexibility for modeling the missing data because of the independent information from the two samples about domain  $ab$ .

We first state conditions for a reduced model to be locally identifiable. Let  $\theta$  denote the  $s$ -vector of parameters of interest; in our case,  $\theta$  would include linearly independent components of  $\mathbf{p}$ ,  $N_{ab}/N$ , and parameters for the missing data mechanism. In the likelihood in (2), the probabilities from the independent multinomial samples are  $Q_{kld}^A$  and  $Q_{kld}^B$ . These probabilities may be written as functions of  $\theta$ , with  $\mathbf{Q}^A(\theta) = (Q_{00a}^A, \dots, Q_{lMab}^A)$  a  $g$ -vector of the nonzero  $Q_{kld}^A$ 's and  $\mathbf{Q}^B(\theta) = (Q_{00b}^B, \dots, Q_{lMab}^B)$  a  $q$ -vector of the nonzero  $Q_{kld}^B$ 's. When all cells in Table 2 and the analogous table for frame  $B$  have nonzero probabilities,  $g = q = 16$ . Let  $\mathbf{D} = (\mathbf{D}'_A, \mathbf{D}'_B)'$  be the derivative matrix of the transformation, with  $\mathbf{D}_{A(\alpha\beta)} = \partial Q_{\alpha}^A / \partial \theta_{\beta}$  and  $\mathbf{D}_{B(\delta\beta)} = \partial Q_{\delta}^B / \partial \theta_{\beta}$  for  $\alpha = 1, \dots, g-1$ ,  $\delta = 1, \dots, q-1$ , and  $\beta = 1, \dots, s$ . Then, using Theorems 3, 4 and 5 in Catchpole and Morgan (1997), the model is locally identifiable if the matrix  $\mathbf{D}$  is of full rank. The proof for the dual frame situation is given in Lu (2007).

In a dual frame survey, we consider two types of models for the missing data. In a Type (1) model, the probabilities of missing time-1 or time-2 information for cell  $(k, l)$  is the same for each domain within a frame, i.e.,  $\phi_{kla}^A = \phi_{klab}^A = \phi_{kla}^B$ ,  $\psi_{kla}^A = \psi_{klab}^A = \psi_{kla}^B$ ,  $\phi_{klb}^B = \phi_{klab}^B = \phi_{klb}^A$  and  $\psi_{klb}^B = \psi_{klab}^B = \psi_{klb}^A$ . In this type of model, we estimate the  $\phi$ 's and  $\psi$ 's separately from each sample. It might be considered when the samples from the two frames are collected using different modes. For example, if the frame  $A$  sample is a mail survey and the frame  $B$  sample is a cell phone survey, one might expect different probabilities of dropout from the two samples.

In a Type (2) model, the probabilities of having missing data are the same in each domain, i.e.,  $\phi_{klab}^A = \phi_{klab}^B = \phi_{klab}$ . This type of model might be considered when nonresponse is expected to be related to the cell membership, and frame membership is thought to have little effect on nonresponse.

For example, if the two surveys have similar types of designs and administrative procedures, a Type (2) model might be appropriate.

For each type of model, we may need to place additional restrictions on the parameters in order to solve the likelihood equations. Following Stasny and Fienberg (1986) the following are possible restrictions:

$$\text{Model 1: } \phi_{kl} = \lambda_{t-1(l)}, \psi_{kl} = \lambda_{t(k)} \quad (3)$$

$$\text{Model 2: } \phi_{kl} = \lambda_{t-1}, \psi_{kl} = \lambda_t$$

$$\text{Model 3: } \phi_{kl} = \lambda_l, \psi_{kl} = \lambda_k$$

$$\text{Model 4: } \phi_{kl} = \lambda_{t-1(l)}, \psi_{kl} = \lambda_t$$

$$\text{Model 5: } \phi_{kl} = \lambda_{t-1}, \psi_{kl} = \lambda_{t(k)}.$$

Under model 1, the probability that an individual is a nonrespondent in a given time period depends on the given time period and the individual's classification in the observed time period. Under model 2, the probability that an individual is a nonrespondent in a given time period depends only on the given time period. Under model 3, the probability that an individual is a nonrespondent in a given time period depends only on the individual's classification in the observed time period. Under model 4, the probability that an individual is a nonrespondent at time 1 depends on that time period and the individual's classification in the observed month, and the probability that an individual is a nonrespondent at time 2 depends only on the time period 2. Under model 5, the probability that an individual is a nonrespondent at time 1 depends only on the time period, and the probability that an individual is a nonrespondent at time 2 depends on the time period and the individual's classification in the observed month. Many other models are possible in addition to these five models for each type. Using the derivative matrices, it is easily shown that Models 1-5 are all identifiable.

In general, we will not have closed form solutions for the parameter estimates and the parameters must be estimated using an iterative method. We use the function 'nlm' in R (www.r-project.org) to calculate parameter estimates; the code is available from the authors.

### 3.3 Estimators from complex samples

When either or both samples are collected with a complex design, using the cell counts directly in the likelihood in (2) will give estimators that are not design-consistent. Skinner and Rao (1996) used a pseudo-maximum likelihood (PML) method to obtain design-consistent estimators in cross-sectional dual frame surveys. They showed that, unlike the estimators of Hartley (1962) and Fuller and



Burmeister (1972), the PML estimators for different response variables used the same set of modified weights and thus were internally consistent.

We propose to study estimators inspired by the PML method for gross flows in dual frame longitudinal complex surveys that allow for missing data at either time period in either sample. The basic idea is to use a working assumption of a multinomial distribution from a finite population to give the form of the estimators and use a design effect to adjust the cell counts to reflect the complex survey design.

In the simple random sampling case,  $x_{kld}^A/n_A$  is a design-consistent estimator of  $Q_{kld}^A$ . To obtain a pseudo-likelihood for general sampling designs, we replace  $x_{kld}^A/n_A$  by  $\hat{X}_{kld}^A/N_A$ , a design-consistent estimator of  $Q_{kld}^A$  under the complex sampling design, in the likelihood (2). Define  $\bar{x}_{kld}^A = \bar{n}_A \hat{X}_{kld}^A/N_A$  and  $\bar{x}_{kld}^B = \bar{n}_B \hat{X}_{kld}^B/N_B$ , where, following Skinner and Rao (1996), we allow  $\bar{n}_A$  and  $\bar{n}_B$  to be arbitrary constants. Note that if  $N_A$  or  $N_B$  is unknown, it may be estimated by  $\hat{N}_A$  or  $\hat{N}_B$  instead.

The pseudo-likelihood has the same form as (2), with  $x_{kld}^A, x_{kld}^B, n_A$  and  $n_B$  replaced by  $\bar{x}_{kld}^A, \bar{x}_{kld}^B, \bar{n}_A$  and  $\bar{n}_B$ , respectively. Iterative procedures are then used to find the pseudo-MLEs of the quantities of interest  $p_{kld}, \phi, \psi$  and  $N_{ab}$ . By the fact that the pseudo-likelihood factors,  $\hat{N}_{ab}$  is found to be the smaller of the roots of

$$\begin{aligned} & [\bar{n}_A + \bar{n}_B] \hat{N}_{ab, PML}^2 \\ & - [\bar{n}_A \hat{N}_B + \bar{n}_B \hat{N}_A + \bar{n}_A \hat{N}_{ab}^A + \bar{n}_B \hat{N}_{ab}^B] \hat{N}_{ab, PML} \\ & + [\bar{n}_A \hat{N}_{ab}^A \hat{N}_B + \bar{n}_B \hat{N}_{ab}^B \hat{N}_A] = 0. \end{aligned} \quad (4)$$

In a complex survey, particularly when clustering is involved, the actual sample sizes  $n_A$  and  $n_B$  do not necessarily reflect the relative amounts of information from the samples. We thus suggest taking  $\bar{n}_A$  and  $\bar{n}_B$  to be the effective sample size for each sample, with  $\bar{n}_A = n_A / (\text{design effect of } S_A)$  and  $\bar{n}_B = n_B / (\text{design effect of } S_B)$ . The design effect of an estimator  $\hat{\mu}$  is the ratio

$$\frac{[V(\hat{\mu}) \text{ from complex survey design}]}{[V(\hat{\mu}) \text{ from SRS of same size}]}$$

The design effect is usually different for different variables. For estimating gross flows, however, the only estimators used from the component surveys are estimated cell counts, and we might expect that in many surveys the design effects for the estimators  $\hat{X}_{kld}^A$  would all be similar, and would also be similar to the design effect of the estimator  $\hat{N}_{ab}^A$ . We thus, as in Skinner and Rao (1996), suggest using the design effect for the estimator  $\hat{N}_{ab}^A$  in determining  $\bar{n}_A$ , and the design effect for the estimator  $\hat{N}_{ab}^B$  in determining  $\bar{n}_B$ . If the design effects of the other variables are indeed identical, then the resulting PMLEs will minimize the variances of the estimated quantities; if they

differ, the PMLEs will not be optimal but they will be consistent and in most situations will be close to the optimal values (Lohr and Rao 2006). If the design effect for  $\hat{N}_{ab}^A$  is unavailable, as would occur, for example, if the survey were poststratified to  $N_{ab}^A$ , then we suggest using a generalized design effect, computed by taking an average or weighted average of design effects from other variables in the survey.

## 4. Properties of the estimators

In this section, we will investigate properties of the estimators. We derive asymptotic variances, discuss jackknife variance estimators, and perform a small simulation study to explore the properties.

### 4.1 Properties

We consider the general case in which stratified multi-stage samples are taken from each frame. The estimators of population totals are the standard Horvitz-Thompson or Hájek estimators from complex surveys. From frame  $A$ , the parameter vector  $\eta_A = [(Q^A)', N_{ab}/N_A]'$  is estimated by  $\hat{\eta}_A = [(\hat{Q}^A)', \hat{N}_{ab}^A/N_A]'$ , where  $\hat{Q}_{kld}^A = \hat{X}_{kld}^A/N_A$ ; similarly,  $\eta_B = [(Q^B)', N_{ab}/N_B]'$  is estimated by  $\hat{\eta}_B = [(\hat{Q}^B)', \hat{N}_{ab}^B/N_B]'$  with  $\hat{Q}_{kld}^B = \hat{X}_{kld}^B/N_B$ .

*Theorem 1:* Let  $\hat{\eta} = (\hat{\eta}_A', \hat{\eta}_B')'$  and  $\eta = (\eta_A', \eta_B')'$ . Assume that the regularity conditions on the inclusion probabilities in Isaki and Fuller (1982) hold for each sample. Let  $\tilde{n}_A$  and  $\tilde{n}_B$  be the number of primary sampling units in frames  $A$  and  $B$ , respectively, and let  $\tilde{n} = \tilde{n}_A + \tilde{n}_B$ . Assume that  $\tilde{n}_A$  and  $\tilde{n}_B$  both increase such that  $\tilde{n}_A/\tilde{n}_B \rightarrow \gamma$  for some  $0 < \gamma < 1$ . Then  $\hat{\eta}$  is consistent for  $\eta$ , and

$$\tilde{n}^{1/2} (\hat{\eta} - \eta) \xrightarrow{d} N(0, \Sigma), \quad (5)$$

where  $\Sigma$  is a block-diagonal matrix with blocks  $\Sigma_A$  and  $\Sigma_B$ ,  $\Sigma_A$  is the asymptotic covariance matrix of  $\tilde{n}_A^{1/2} \hat{\eta}_A$  and  $\Sigma_B$  is the asymptotic covariance matrix of  $\tilde{n}_B^{1/2} \hat{\eta}_B$ . If, in addition, it is assumed that  $N_{ab}/N \rightarrow \kappa$  for some  $0 < \kappa < 1$  and that the model is identifiable, then  $\hat{\theta}$  is consistent for  $\theta$ , where  $\theta$ , the parameter of interest, consists of components of  $\mathbf{p}, N_{ab}/N, \phi$  and  $\psi$ , and  $\hat{\theta}$  is the pseudo-maximum likelihood estimator of  $\theta$ . Furthermore,  $\tilde{n}^{1/2} (\hat{\theta} - \theta)$  is asymptotically normal with mean 0 and asymptotic variance  $\mathbf{H}_A \Sigma_A \mathbf{H}_A' + \mathbf{H}_B \Sigma_B \mathbf{H}_B'$ , where  $\mathbf{H}_F$  is the derivative matrix of the function  $\theta$  with respect to the parameters  $\eta_F$  for frames  $F \in \{A, B\}$ .

*Proof.* With gross flows, observed values of all variables are 0 or 1. Thus the boundedness conditions in Lemmas 1 and 2 of Isaki and Fuller (1982) are met, and the estimators of frame  $A$  are consistent and asymptotically normal with

$$\tilde{n}_A^{1/2} (\hat{\eta}_A - \eta_A) \xrightarrow{d} N[0, (\gamma/(1 + \gamma)) \Sigma_A].$$

The same argument applies to give consistency and asymptotic normality for the vector of estimators from frame  $B$ , with

$$\tilde{n}_B^{1/2}(\hat{\boldsymbol{\eta}}_B - \boldsymbol{\eta}_B) \xrightarrow{d} N[0, (1 - (\gamma/(1 + \gamma))) \boldsymbol{\Sigma}_B].$$

Combining these two asymptotic results, and using the independence of the sampling designs along with Slutsky’s theorem, gives (5). The limiting distribution of  $\tilde{n}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$  follows by the delta method, since the parameters in  $\boldsymbol{\theta}$  are all twice continuously differentiable functions of those in  $\boldsymbol{\eta}$ . Since the parameter estimators cannot always be defined explicitly as a function of other statistics from the sample, we may derive the matrices  $\mathbf{H}_A$  and  $\mathbf{H}_B$  by linearizing the score equations (Binder 1983). The assumption that  $N_{ab}/N \rightarrow \kappa \in (0, 1)$  guarantees that the linearization is well-defined.

Theorem 1 shows that linearization can be used to estimate the variances of parameters of interest. In many situations, however, the matrices  $\mathbf{H}_A$  and  $\mathbf{H}_B$  are high-dimensional and the linearized variance estimators have complex form. A practical way to estimate the variances of the estimators is to use the jackknife estimator proposed by Lohr and Rao (2000). Under the regularity conditions in their Theorem 4, the jackknife and linearization variance estimators are asymptotically equivalent. The form of the jackknife variance estimator is  $v_{JK}(\hat{\boldsymbol{\theta}}) = v_A(\hat{\boldsymbol{\theta}}) + v_B(\hat{\boldsymbol{\theta}})$ , where  $v_A$  is a jackknife estimator obtained by deleting one primary sampling unit at a time from frame  $A$  while using the full data set for frame  $B$ , and  $v_B$  is a jackknife estimator obtained by deleting one primary sampling unit at a time from frame  $B$  while using the full data set for frame  $A$ .

4.2 Simulation study

Theorem 1 shows that the dual frame estimators are consistent for the corresponding population quantities under the modeled missing data mechanism. We performed a small simulation study to investigate properties for moderate sample sizes with overlapping frames. We generated the data following the simulation study in Skinner and Rao (1996), with  $\gamma_a = N_a/N$  and  $\gamma_b = N_b/N$ . A cluster sample from frame  $A$  was generated with  $\tilde{n}_A$  psus and  $m$  observations in each psu, and a simple random sample of  $n_B$  observations was generated for frame  $B$ . We generated the clustered binary responses for the sample from frame  $A$  by generating correlated multivariate normal random vectors and then using the probit function to convert the continuous responses to binary responses.

After generating the sample, we calculated the estimators of the probabilities of the union of frame  $A$  and frame  $B$ , average of the absolute value of the bias and empirical mean

squared error (EMSE) under different settings. The EMSE of a given estimator,  $\hat{Y}$  is calculated as:

$$EMSE = \frac{1}{R} \sum_{r=1}^R (\hat{Y}_r - Y)^2, \tag{6}$$

where  $\hat{Y}_r$  is the value of  $\hat{Y}$  for the  $r^{th}$  simulation run. In our simulation study, we used  $R = 100$ .

The simulation study was performed with factors: (1)  $\gamma_a$ : 0.2 or 0.4, (2)  $\gamma_b$ : 0.2 or 0.4, (3) clustering parameter  $\rho$ : 0.3, (4) missing data mechanism: the probability that an individual is a nonrespondent in a given month depends on the time period and the individual’s classification in the observed period; or missing completely at random, (5) amount of missing data: close to 10% or close to 20%, (6) sample sizes:  $\tilde{n}_A$ : 10, 100 or 500;  $m$ : 5,  $n_B$ : 100, 1,000 or 5,000. All runs used probability parameters  $\mathbf{p}_a$ : (0.3, 0.1, 0.2, 0.4),  $\mathbf{p}_{ab}$ : (0.3, 0.1, 0.1, 0.5), and  $\mathbf{p}_b$ : (0.4, 0.1, 0.1, 0.4). Table 3 shows the results of the simulation study with missing data generated under Model 1 and fitted with both Model 1 and the model using complete records only.

**Table 3**  
Results from the simulation study for missing data generated under Model 1. Case (1) fits the correct model: Model 1; Case (2) uses complete records only. Bias is the average absolute bias for the population gross flow proportions  $p_{kl}$ ; EMSE is the average empirical mean squared error for the  $p_{kl}$ ; the proportions used to generate the missing data are  $\lambda_{(t-1)0} = 0.141$ ,  $\lambda_{(t-1)1} = 0.070$ ,  $\lambda_{(t)0} = 0.137$  and  $\lambda_{(t)1} = 0.068$ . Here,  $\tilde{n}_A$  is the number of psus in sample  $A$  with psu size 5 and  $n_B$  is the number of elements in sample  $B$

$\tilde{n}_A$	$n_B$		$P_{00}$	$P_{01}$	$P_{10}$	$P_{11}$
10	100	Estimator	0.311	0.120	0.149	0.420
		Bias	0.040	0.029	0.029	0.040
		EMSE	0.002	0.001	0.001	0.002
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{t(1)}$
		Estimator	0.159	0.095	0.146	0.094
		EMSE	0.001	0.001	0.002	0.001
10	100	Estimator	0.286	0.120	0.146	0.448
		Bias	0.048	0.029	0.029	0.041
		EMSE	0.004	0.001	0.001	0.002
100	1,000	Estimator	0.321	0.092	0.138	0.449
		Bias	0.015	0.011	0.009	0.015
		EMSE	3.337e-04	1.798e-04	1.418e-04	3.256e-04
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{t(1)}$
		Estimator	0.145	0.074	0.123	0.068
		EMSE	2.642e-04	9.389e-05	3.917e-04	8.206e-05
100	1,000	Estimator	0.293	0.092	0.135	0.480
		Bias	0.0280	0.011	0.010	0.040
		EMSE	0.001	1.839e-04	1.711e-04	0.002
500	5,000	Estimator	0.321	0.093	0.135	0.452
		Bias	0.006	0.008	0.007	0.012
		EMSE	4.960e-05	7.162e-05	6.381e-05	1.857e-04
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{t(1)}$
		Estimator	0.140	0.071	0.123	0.064
		EMSE	4.466e-05	1.818e-05	2.288e-04	3.545e-05
500	5,000	Estimator	0.292	0.092	0.132	0.483
		Bias	0.028	0.008	0.008	0.043
		EMSE	8.265e-04	7.642e-05	9.571e-05	1.906e-03



When data are missing at random, all models give estimators of the gross flow proportions  $p_{kl}$  that are approximately unbiased so we do not report the results here. From Table 3, both the correct model and the analysis of complete records only produce biased estimators of the  $p_{kl}$ 's. With larger sample sizes, however, the bias persists in the analysis that uses complete records only, while it diminishes when Model 1 is fit. This example has relatively small probabilities of missing data. With larger amounts of missing data, the contrast between the estimators is more pronounced.

## 5. Application

In this section, we apply our results to data from the Survey of Income and Program Participation (SIPP) and the Current Population Survey (CPS) within Arizona. Both CPS and SIPP are longitudinal stratified multistage panel surveys. We treat SIPP and CPS as a dual frame survey with the same target population: the Arizona population 18 years old to 64 years old. Using information from both surveys, we want to model the transition probabilities of employment status changes from January 2001 to January 2002 of people between 18 years old and 64 years old. Note that, strictly speaking, these two surveys are not designed as a dual frame survey. They use different questions for the labor force variables. Although we recoded the variables according to the labor force definitions in CPS, it is possible that these different question wordings and orderings produce bias when combining the information. We use this as an example because a real longitudinal dual frame data is not available. Nevertheless, the example shows the potential gains in efficiency by combining the information from two surveys in estimating gross flows.

Both surveys have target population the noninstitutionalized civilian population of the United States. We consider a subset of the population: the population in the labor force from 18 years old to 64 years old. So  $N_A = N_B = N_{ab}$  and the estimation problem is a special case of the theory given in Section 3. The longitudinal file for the 2001 and 2002 SIPP (Westat 2001) uses one panel. We merged Wave 1 (where January 2001 records are stored), Wave 4 (where January 2002 records are stored) and the longitudinal weight file, in which the weights are adjusted to sum to the population count. Since the longitudinal panel weights have been adjusted for the nonresponse, we consider this as a no missing data case. The resulted weighted gross flow table from SIPP is given in Table 4.

For the CPS, the rotation group design introduces partially classified data. January 2001 and January 2002 have 50 percent of the sample in common. We use these 50% of the data together with the partially classified data to

perform the analysis. The weight variable we use is a cross-sectional weight with cross-sectional nonresponse and calibration adjustments (United States Census Bureau 2006). For individuals present in the survey for only one of the years, we use the weight from that year. For persons present in both Jan 2001 and Jan 2002, we use the average of the two weights. The rule that we chose the average of the two weights is to minimize the variance of the composite estimator. The population group we used is the 18-64 age group, and we excluded persons who were not in that category during both years. The weighted gross flow table from CPS is in Table 5.

**Table 4**  
Gross flow table for SIPP, in Arizona

		Jan 2002	
		Employed	Unemployed
January 2001	Employed	2,491,029	73,204
	Unemployed	30,698	30,160
		2,625,091	

**Table 5**  
Gross flow table for CPS, in Arizona

		January 2002		
		Employed	Unemployed	Missing
January 2001	Employed	1,129,656	38,848	689,497
	Unemployed	41,586	8,211	36,041
	Missing	606,549	57,549	
		2,607,937		

Since SIPP is considered as a no missing data case, we assumed  $\phi_{kl} = \psi_{kl} = 0$  and use a Type 1 model in the data analysis. We adjusted each weight in the CPS data by the factor  $2,625,091/2,607,937$  to reach a single population total between the two time periods and a single population total between the two surveys. The number of observations in SIPP (frame A) after combining January 2001 and January 2002 are 551 and the design effect for unemployment is about 1.76, so  $\bar{n}_A = 551/1.76 = 313$ . The design effect for unemployment in CPS (frame B) is about 1.229, so  $\bar{n}_B = 1,020/1.229 = 830$ . Because the likelihood factors, the estimated parameters of probabilities from the five models (3) are all the same. We list the estimated probabilities and the standard errors from SIPP, CPS and data combining these two surveys in Table 6.

**Table 6**  
Estimated transition probabilities using SIPP, CPS, and the dual frame method with SIPP and CPS. Standard errors are given in parentheses

	$P_{00}$	$P_{01}$	$P_{10}$	$P_{11}$
SIPP	0.9489 (0.0124)	0.0279 (0.0093)	0.0117 (0.0061)	0.0115 (0.0060)
CPS	0.9088 (0.0100)	0.0454 (0.0072)	0.0353 (0.0064)	0.0106 (0.0035)
SIPP and CPS	0.9230 (0.0080)	0.0381 (0.0058)	0.0262 (0.0050)	0.0127 (0.0030)

Due to confidentiality issues, no clustering information is available in the CPS public-use data sets. We used a product of the published design effect and the variance from multinomial sampling to estimate the variances from both SIPP and CPS data. The result from Theorem 1 was applied to estimate the variances of  $\hat{p}_{kl}$  for  $k, l = 0, 1$ . In this special situation, the variance estimate from the combination of the two data sets is reduced to  $(\bar{n}_A/(\bar{n}_A + \bar{n}_B))^2 V_A + (\bar{n}_B/(\bar{n}_A + \bar{n}_B))^2 V_B$ , where  $V_A$  denotes the variance estimate from SIPP data and  $V_B$  denotes the variance estimate from CPS data. Table 6 shows that the standard errors are reduced by using the dual frame method.

We also performed goodness-of-fit tests, developed in Lu (2007), for the five models in (3). The parameter estimates from the five models and results from the goodness-of-fit tests, are listed in Table 7. All five models fit the data well, so we recommend adopting the simplest model, Model 3, for the data.

**Table 7**  
Estimated parameters and results of goodness of fit tests

	Estimated Parameters				df	Corrected $G^2$	$p$ -value
Model 1	$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{t(1)}$	3	3.03	0.39
	0.246	0.395	0.277	0.302			
Model 2	$\lambda_{t-1}$	$\lambda_t$			5	8.58	0.12
	0.255	0.278					
Model 3	$\lambda_0$	$\lambda_1$			5	6.61	0.25
	0.262	0.353					
Model 4	$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_t$		4	4.10	0.39
	0.246	0.397	0.278				
Model 5	$\lambda_{t-1}$	$\lambda_{t(0)}$	$\lambda_{t(1)}$		4	6.74	0.15
	0.255	0.277	0.313				

With the limited information available on the public-use data sets, we used simple weight adjustments to make the estimated population counts consistent with known totals. The SIPP and CPS weights in the data sets have already been calibrated and adjusted for nonresponse, so that the models for missing data mostly reflect the rotating panel design rather than attrition due to moving and other activities that might be related to employment status.

Future research on these models might include using different weighting adjustments for the longitudinal surveys. In addition, different parameters could be used to distinguish observations that have partial classification because of the rotating panel design, and observations that have partial classification because of nonresponse. To do so, we could introduce a Markov Chain model similar to the one proposed by Stasny (1987). In the complete data model, individuals are allocated to the table according to a single multinomial distribution. At the second step of the process, which is also unobserved, each individual may be chosen to either rotate out of the sample after the interview for month

$t - 1$  or rotate into the sample before the month  $t$  interview according to the sampling plan. Finally, in the third step of the process, each remaining individual may either lose its row classification or lose its column classification by other reasons. Using this model, we can model the nonresponse at both times (*i.e.*, lose both the row and the column classifications).

## 6. Conclusions

In this article, we developed statistical methods for estimating gross flows from dual frame surveys. These methods are necessary to estimate changes in poverty status or employment status over time. We developed pseudo-maximum likelihood estimators that use the dual frame structure and the properties of the two survey designs. Our models also account for effects of missing data when an individual drops out of the survey or when a rotation panel design is used, so they allow full use of partial information that may be provided by some households. We use a jackknife method to estimate the variance of estimators and examine the properties of the estimators. The results have been applied to real datasets.

In this paper, the categories of the gross flow tables are defined independently from the sample outcomes. It is also possible to define the categories based on values that depend on the sample. For example, in social surveys, the poverty line might be defined using a percentile from the sample and the categories defined as “Below the poverty line” and “Above the poverty line.” Methods from this paper can be used to estimate gross flows if the category definitions depend on the sample, but the variance estimators need to account for the effect of estimating the category boundaries.

Although the results in this paper are for dual frame surveys, the methods are general and could be extended to more than two surveys using PML estimators developed in Lohr and Rao (2006). As the number of frames increases, however, so does the complexity of possible missing data mechanisms. Misclassification error may also be more prevalent with a larger number of frames.

Our research is done in the context of survey sampling, but it also applies to other settings in which data could be combined from two independent sources. As it becomes increasingly difficult for a single survey to cover the entire population of interest, we believe these methods for estimating gross flows can provide better coverage of the population with less expense. They also allow for supplementing a general population survey with surveys of specific subpopulations of interest.

## Acknowledgements

This research was partially supported by the National Science Foundation under grants SES-0604373 and



DLS-0909630. The authors thank the associate editor and referees for their insightful and helpful comments.

## References

- Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Blair, E., and Blair, J. (2006). Dual frame web-telephone sampling for rare groups. *Journal of Official Statistics*, 22, 211-220.
- Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association*, 63, 542-551.
- Catchpole, E.A., and Morgan, B.J.T. (1997). Detecting parameter redundancy. *Biometrika*, 84, 187-196.
- Chambers, R.L., Woyzbun, L. and Pillig, R. (1988). Maximum likelihood estimation of gross flows. *Australian Journal of Statistics*, 30, 149-162.
- Chen, T., and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.
- Fuller, W.A., and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. In *Proceedings of the Social Statistics Section*, American Statistical Association, 245-249.
- Hartley, H.O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.
- Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, Series C, 36, 99-118.
- Heeringa, S.G. (1995). Technical description of the assets and health dynamics (ahead) survey sample design. Technical Paper, Institute for Social Research, University of Michigan, [hrsonline.isr.umich.edu/docs/userg/AHDSAMP.pdf](http://hrsonline.isr.umich.edu/docs/userg/AHDSAMP.pdf).
- Hocking, R.R., and Oxspring, H.H. (1971). Maximum likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, 66, 65-70.
- Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.
- Lu and Lohr: Gross flow estimation in dual frame surveys
- Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A*, 149, 65-82.
- Lavallée, P. (2007). *Indirect Sampling*. New York: Springer-Verlag.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
- Lohr, S.L., and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.
- Lohr, S.L., and Rao, J.N.K. (2006). Estimation in Multiple-frame Surveys. *Journal of the American Statistical Association*, 101, 1019-1030.
- Lu, Y. (2007). Longitudinal estimation in dual frame surveys. *Ph.D Dissertation, Arizona State University*.
- Pfeffermann, D., Skinner, C. and Humphreys, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society, Series A*, 161, 13-32.
- Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Stasny, E.A. (1984). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 25-40.
- Stasny, E.A. (1987). Some Markov-Chain models for nonresponse in estimating gross labor force flows. *Journal of Official Statistics*, 4, 359-73.
- Stasny, E.A., and Fienberg, S.E. (1986). Estimating gross flows using panel data with nonresponse: An example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.
- United States Census Bureau (2006). Current Population Survey: Design and Methodology. Technical Paper 66, U.S. Census Bureau, Washington, DC.
- Verma, V., Betti, G. and Ghellini, G. (2007). Cross-sectional and longitudinal weighting in a rotational household panel: application to EU-SILC. *Statistics in Transition*, 8, 5-50.
- Westat (2001). Survey of Income and Program Participation Users' Guide (Supplement to the Technical Documentation). Technical report, Washington, DC.

# Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling

Qixuan Chen, Michael R. Elliott and Roderick J.A. Little<sup>1</sup>

## Abstract

We propose a Bayesian Penalized Spline Predictive (BPSP) estimator for a finite population proportion in an unequal probability sampling setting. This new method allows the probabilities of inclusion to be directly incorporated into the estimation of a population proportion, using a probit regression of the binary outcome on the penalized spline of the inclusion probabilities. The posterior predictive distribution of the population proportion is obtained using Gibbs sampling. The advantages of the BPSP estimator over the Hájek (HK), Generalized Regression (GR), and parametric model-based prediction estimators are demonstrated by simulation studies and a real example in tax auditing. Simulation studies show that the BPSP estimator is more efficient, and its 95% credible interval provides better confidence coverage with shorter average width than the HK and GR estimators, especially when the population proportion is close to zero or one or when the sample is small. Compared to linear model-based predictive estimators, the BPSP estimators are robust to model misspecification and influential observations in the sample.

**Key Words:** Bayesian analysis; Binary data; Penalized spline regression; Probability proportional to size; Survey samples.

## 1. Introduction

Unequal probability sampling designs are commonly employed in data collection by science and government. Perhaps the simplest unequal probability design is stratified sampling, which samples units from different strata with different inclusion probabilities. Another important form of unequal probability sampling is probability-proportional-to-size (pps) sampling, in which the inclusion probability is proportional to the value of a size variable measured for all population units.

An unequal probability sampling design such as pps sampling is often used for efficient estimation of population means of continuous variables, for which the variance increases with size of unit. However, inferences about discrete variables are often also of interest in a multipurpose survey (*e.g.*, Lehtonen and Veijanen 1998, Lehtonen, Särndal and Veijanen 2005). In this paper, we focus on methods of inference for finite population proportions from unequal probability sampling designs, based on an auxiliary variable measured for all the units in the population. We use pps sampling as a specific design to illustrate and assess our methods.

The inclusion probabilities play important and somewhat different roles in design-based and model-based inference from unequal probability survey samples (Smith 1976, 1994; Kish 1995; Little 2004). In design-based inference, survey variables are fixed, and inference is based on the distribution of the sample inclusion indicators; the standard design-based approaches to estimation such as the Horvitz-Thompson

(HT) estimator (1952) and its extensions weight sampled units by the inverse of their inclusion probabilities. These estimators are design consistent (Isaki and Fuller 1982) and provide reliable inferences in large samples without the need for modeling assumptions. However, these estimators are potentially very inefficient, as illustrated in Basu's (1971) famous elephant example. Also, variance estimation is cumbersome because it requires second-order inclusion probabilities. Corresponding confidence intervals are based on asymptotic theory, and may deviate from nominal levels for moderate or small sample sizes.

Model-based inference predicts values of survey variables in the non-sampled units by including the inclusion probabilities as covariates in the prediction model (Little 2004). Model-based prediction estimators are consistent and efficient under the assumed model, but are subject to bias when the underlying model is misspecified. This limitation motivates the development of flexible statistical models that are more robust to model misspecification. For continuous survey data, Zheng and Little (2003) estimated the finite population total using a nonparametric regression on a penalized spline (*p*-spline) of the inclusion probabilities. We propose here Bayesian *P*-Spline Predictive (BPSP) estimators that are suitable for a binary, as opposed to continuous, outcome. We adopt a Bayesian approach to inference for this model, since Bayesian methods often yield better inference for small sample problems, and are conveniently implemented for our proposed model via the Gibbs' sampler. In this approach, auxiliary variables other than the inclusion probability can also be included in the model, but

1. Qixuan Chen is Assistant Professor, Department of Biostatistics, Columbia University, 722 West 168 Street, New York, NY 10032. E-mail: qc2138@columbia.edu; Michael R. Elliott is Associate Professor and Roderick J.A. Little is Professor, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: mreliot@umich.edu and rlittle@umich.edu.



the inclusion probability is singled out since modeling of this variable is prone to model misspecification.

We compare the performance of BPSP estimators with Hájek (HK, Horvitz-Thompson-type) estimators and with Generalized Regression (GR) estimators for a binary outcome proposed by Lehtonen and Veijanen (1998). The GR approach is a popular model-assisted modification of the design-based estimators that combines predictions from a model with design-weighted model residuals (Montanari 1998), to yield estimates that are approximately design unbiased.

Zheng and Little (2003; 2005) compared HT,  $p$ -spline prediction, and GR estimates of the total of a continuous survey variable by simulation. They found that  $p$ -spline model-based estimators had better root mean squared error than the other methods, and with jackknife standard errors providing superior confidence coverage to HT or GR inferences. We conduct similar comparisons for inference about a population proportion for a binary outcome, and show similar advantages for our BPSP estimator over the HK and GR alternatives.

## 2. Design-based estimator

Suppose that we have a finite population consisting of  $N$  identifiable units. Let  $Y$  be the binary survey variable of interest and  $p = N^{-1} \sum_{i=1}^N Y_i$  be the proportion of the population for which  $Y = 1$ . Let  $\pi_i$  denote the probability of inclusion for unit  $i$ , which is assumed to be known for all units in the finite population before a sample is drawn. An unequal probability random sample  $s$  with elements  $y_1, \dots, y_n$  is then drawn from the finite population according to the inclusion probabilities  $\pi_1, \dots, \pi_N$ . The design-based HK estimator in the discussion of Basu (1971) is defined as

$$\hat{p}_{\text{HK}} = \frac{\sum_{i \in s} y_i / \pi_i}{\sum_{i \in s} 1 / \pi_i}. \quad (1)$$

The variance for  $\hat{p}_{\text{HK}}$  can be estimated via linearization of the Yates-Grundy estimator (1953) of totals,

$$\hat{V}_{\text{YG}}(\hat{p}_{\text{HK}}) = \left( \sum_{k \in s} 1 / \pi_k \right)^{-2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \left( \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left( \frac{y_i - \hat{p}_{\text{HK}}}{\pi_i} - \frac{y_j - \hat{p}_{\text{HK}}}{\pi_j} \right)^2. \quad (2)$$

The Yates-Grundy variance estimator requires pairwise inclusion probabilities. When the pairwise inclusion probabilities are not available, as in our simulations, the approximate formula proposed by Hartley and Rao (1962),

$$\pi_{ij} \approx \frac{n-1}{n} \pi_i \pi_j + \frac{n-1}{n^2} (\pi_i^2 \pi_j + \pi_i \pi_j^2) - \frac{n-1}{n^3} \pi_i \pi_j \sum_{k=1}^N \pi_k^2,$$

has frequently been used. An approximate  $1 - \alpha$  level confidence interval for the population proportion  $\hat{p}_{\text{HK}}$  is then obtained based on the normal approximation.

## 3. Bayesian $P$ -Spline Predictive (BPSP) estimator

Royall (1970) argued for the use of models for finite-population descriptive inferences by predicting the unobserved values based on models, since model-based inferences should be more efficient than design-based inferences. To model the relationship between the binary outcome  $Y$  and the continuous inclusion probability  $\pi$ , we need to fit a binary regression of  $Y$  on  $\pi$ . Parametric binary regressions, such as the linear or quadratic logistic or probit model, may not be adequate in fitting the data. One solution for this problem of inflexibility is to fit a binary regression on a spline of  $\pi$  by adding some knots. However, too many knots may result in the roughness of model fit. One way to overcome this problem is to retain all of the knots but to constrain their influence, by fitting a binary  $p$ -spline regression model.

Common methods for modeling a binary outcome are logistic and probit regressions, and they generally give similar results. We choose to adopt probit models in our study for computational convenience. The probit regression model for binary outcomes has an underlying truncated normal regression structure on latent continuous data. If the latent continuous data are known, the parameters in binary  $p$ -spline regression models can be estimated using standard approaches for normal  $p$ -spline regression models. In a Bayesian context, the posterior distribution of parameters in the probit  $p$ -spline model can be computed using Gibbs sampling (Albert and Chib 1993; Ruppert, Wand and Carroll 2003, chapter 16). In contrast, the logistic  $p$ -spline regression model requires a more complicated computation procedure such as the Metropolis-Hastings algorithm. The computational advantage makes the probit link function more desirable than the logit link function in Bayesian binary  $p$ -spline regression models.

There are various types of  $p$ -splines. When applying  $p$ -splines, we need to make choices on the degree and knot locations, and the basis functions used to present the model. We choose to use the truncated polynomial  $p$ -splines because they are simple and intuitive. More numerically stable estimators can be obtained using  $B$ -splines via orthogonalizing the truncated power bases (Eilers and Marx

1996). The probit truncated polynomial  $p$ -spline regression model has a generalized linear mixed model representation,

$$\Phi^{-1}(E(y_i | \beta, b, \pi_i)) = \beta_0 + \sum_{k=1}^p \beta_k \pi_i^k + \sum_{l=1}^m b_l (\pi_i - k_l)_+^p \quad (3)$$

$$b_l \sim N(0, \tau^2)$$

$$l = 1, \dots, m; i = 1, \dots, n,$$

where  $\Phi^{-1}(\cdot)$  denote the inverse CDF of a standard normal distribution, and the constants  $k_1 < \dots < k_m$  are  $m$  selected fixed knots. A function such as  $(\pi_i - k)_+^p$  is called a truncated polynomial spline basis function with power  $p$ , where  $(u)_+^p$  is equal to  $\{u \times I(u \geq 0)\}^p$  for any real number  $u$ . Since the truncated polynomial spline basis function has  $p - 1$  continuous derivatives, higher values of  $p$  lead to smoother spline functions. By specifying a normal distribution for  $b$ , the influence of the  $m$  knots is constrained in Model (3), which is equivalent to smooth the splines via the penalized likelihood.

The parameters in Model (3) can be estimated using generalized linear mixed model methods. An alternative Bayesian approach that simplifies computation is to assume weak prior and hyperprior distributions and use Gibbs sampling to obtain draws from the posterior distributions of the parameters as follow: the probit regression model for binary responses has an underlying normal regression structure on latent continuous data; if the latent data are known, the posterior distribution of the parameters can be computed using standard results for normal regression models; and given the posterior distribution of the parameters, the latent continuous data can be simulated from a suitable truncated normal distribution. (Ruppert *et al.* 2003, page 290) The detailed algorithm of Gibbs sampling is in the Appendix. In addition, the Bayesian inference for  $p$ -spline regression can also been implemented using WinBUGS, the standard Bayesian analysis software (Crainiceanu, Ruppert and Wand 2005).

The posterior distribution of the population proportion is simulated by generating a large number  $D$  of draws and using the predictive estimator form  $\hat{p}_{PR}^{(d)} = N^{-1}(\sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j^{(d)})$ , where  $\hat{y}_j^{(d)}$  is a draw from the posterior predictive distribution of the  $j^{\text{th}}$  non-sampled unit of the binary outcome. The average of these draws simulates the Bayesian  $P$ -Spline Predictive (BPSP) estimator of the finite population proportion, and is denoted as  $\hat{p}_{BPSP}$ , where

$$\hat{p}_{BPSP} = D^{-1} \sum_{d=1}^D \hat{p}_{PR}^{(d)}. \quad (4)$$

The Bayesian analog of a  $100 \times (1 - \alpha)\%$  confidence interval for the population proportion is a  $100 \times (1 - \alpha)\%$

credible interval, which can be formed in a number of different ways. We split the tail area  $\alpha$  equally between the upper and lower endpoints in the simulations.

Firth and Bennett (1998) showed that any parametric logistic regression model containing an intercept term and the inverse of inclusion probabilities as a covariate, fitted by ordinary, unweighted maximum likelihood, was “internally bias calibrated” (IBC) for population proportions, and thus yields design consistency. This property is also true for logistic truncated polynomial  $p$ -spline regression models on the inverse of inclusion probabilities, fitted via penalized likelihood. With the probit link function used instead of the logit link function and fitted via Markov chain Monte Carlo algorithm instead of maximum penalized likelihood, the BPSP estimator may no longer have the IBC property. However, the similarity between the probit model and the logistic model implies that the predictive estimator based on a probit  $p$ -spline regression model is approximately design-consistent. We believe that obtaining efficient estimates with close to nominal confidence coverage in finite samples is more important than exact design consistency.

#### 4. Generalized Regression (GR) estimator

For the estimation of class frequencies of a discrete response variable, Lehtonen and Veijanen (1998) proposed a GR estimator  $\hat{t}_{GR}$  of the total, which combines the predicted values  $\hat{y}_i = \hat{\Pr}(Y_i = 1 | \pi_i)$  based on a suitable model and the HT estimator for the residuals  $r_i = y_i - \hat{y}_i$  of the sampled units,

$$\hat{t}_{GR} = \sum_{i=1}^N \hat{y}_i + \sum_{i \in s} r_i / \pi_i. \quad (5)$$

The GR estimator in Equation (5) is then used in constructing an estimator for population proportions by dividing by the known population size  $N$  (Duchesne 2003),

$$\hat{p}_{GR-1} = \frac{1}{N} \left( \sum_{i=1}^N \hat{y}_i + \sum_{i \in s} r_i / \pi_i \right). \quad (6)$$

We also consider here another version of the GR estimator for the estimation of finite population proportions, in which the denominator of the bias calibration term for the residuals  $r_i$  is the estimated population size  $\sum_{i \in s} 1 / \pi_i$ ,

$$\hat{p}_{GR-2} = \frac{1}{N} \sum_{i=1}^N \hat{y}_i + \left( \sum_{i \in s} r_i / \pi_i \right) \left( \sum_{i \in s} 1 / \pi_i \right)^{-1}. \quad (7)$$

For the variance estimate of (6), we use the variance estimator of the estimated total of a discrete response variable, given by Lehtonen and Veijanen (1998), divided by  $N^2$ . For the variance estimate of (7), we apply the



Taylor linearization technique (Särndal, Swensson and Wretman 1992, page 182). These two variance estimators are shown in equations (8) and (9), respectively.

$$\hat{V}(\hat{P}_{GR\_1}) = \frac{1}{N^2} \sum_{k \in S} \sum_{l \in S} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{r_k}{\pi_k} \frac{r_l}{\pi_l}, \quad (8)$$

$$\hat{V}(\hat{P}_{GR\_2}) = \left( \sum_{i \in S} 1/\pi_i \right)^{-2} \sum_{k \in S} \sum_{l \in S} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}, \quad (9)$$

where  $e_k = r_k - (\sum_{i \in S} r_i / \pi_i) (\sum_{i \in S} 1/\pi_i)^{-1}$ . These variance estimators also require pairwise inclusion probabilities, which can be approximated by the method of Hartley and Rao (1962).

However, the Hartley and Rao approximation may lead to bias in the variance estimator. Thus, we also consider the jackknife method for variance estimation (Shao and Wu 1989). The sample is stratified into  $n/G$  strata each of size  $G$  with similar values of inclusion probabilities, and the  $G$  subgroups are then constructed by selecting one element at a time from each stratum without replacement (Zheng and Little 2005). Let  $\hat{p}_{(g)}$  be the same GR estimators in (6) and (7) calculated from the reduced sample without the elements in the  $g^{\text{th}}$  subgroup, and let  $\bar{p}$  be the average of the  $G$  estimators based on the  $G$  reduced samples. The jackknife variance estimator of  $\hat{p}_{GR}$  is

$$\hat{V}_{\text{jackknife}}(\hat{p}_{GR}) = \frac{G-1}{G} \sum_{g=1}^G (\hat{p}_{(g)} - \bar{p})^2. \quad (10)$$

A design-weighted logistic regression model on other covariates was used as the assisting model to predict  $\hat{y}_i$  in the GR estimators for binary outcomes (Lehtonen and Veijanen 1998; Lehtonen *et al.* 2005). Since our interest here is in comparisons of GR estimators with the BPSP estimator, we apply the estimators (6) and (7) with linear probit regression models and probit  $p$ -spline models, as described in detail in Section 5. For the GR estimator using a linear probit model as the assisting model, we use the inclusion probability as a covariate as well as a weight in our simulations.

## 5. Simulation study

### 5.1 Design of the simulation study

Simulation studies are conducted to study the performance of the BPSP estimator compared with the HK estimator, the GR estimators, and the linear model-based predictive estimators for a variety of populations in pps sampling. We present the simulation results for the following six estimators:

- HK, the Hájek estimator defined by equation (1).
- LR, predictive estimator of the form  $\hat{p}_{LR} = N^{-1} (\sum_{i \in S} y_i + \sum_{j \notin S} \hat{y}_j^{\text{LR}})$  with prediction  $\hat{y}_j^{\text{LR}}$  obtained with the maximum likelihood predictions from the linear logistic regression model containing a constant term and the reciprocal inclusion probability as the covariate. LR has the IBC property, and hence is design-consistent. LR is exactly the same as its GR estimator in equation (6).
- PR, predictive estimator of the form  $\hat{p}_{PR} = N^{-1} (\sum_{i \in S} y_i + \sum_{j \notin S} \hat{y}_j^{\text{PR}})$  with prediction  $\hat{y}_j^{\text{PR}}$  from the Bayesian linear probit model containing an intercept term and the inclusion probability as the covariate.
- PR\_GR, the GR estimator in equation (7), where  $\hat{y}_i$  is the prediction for unit  $i$  with unknown parameters replaced by weighted maximum likelihood estimates from the probit model with a constant term and the inclusion probability as the covariate.
- BPSP, the BPSP estimator defined by equation (4) with  $p = 1$  and inverse-gamma prior distribution for  $\tau^2$  and using 15 knots.
- BPSP\_GR, the GR estimator in equation (7), where  $\hat{y}_i$  is the posterior mean of  $\Pr(Y_i = 1 | \pi_i)$  from the BPSP model.

We only report the simulation results based on the linear splines for the BPSP estimator, since simulations not shown here suggest that linear splines perform as well as quadratic splines or cubic splines in all the simulation scenarios. We choose two fixed numbers of knots (15 or 30), and place knots at evenly spaced sample percentiles. The choices of knots work well and a number of 15 knots is good enough to catch the curvatures in our simulations. In addition, the GR estimators in (6) perform similarly to the estimators in (7); some differences between these estimators emerge in the real application in Section 6, leading us to prefer (7) over (6).

We simulated two artificial populations of size 2,000, using two different distributions, with sampling rates of 5% and 10%, where the size variable takes the consecutive integer values 71, 72, ..., 2,070. The inclusion probabilities in the population were then calculated as proportional to the size variable, with the maximum value about 30 times the minimum values.

Continuous data  $Z$  were first generated from normal distributions with mean structure  $f(\pi)$  and constant error variance 0.04. Two different mean structures  $f(\pi)$  were simulated: a linearly increasing function (LINUP)  $f(\pi_i) = k_1 \pi_i$  and an exponential function (EXP)  $f(\pi_i) = \exp(-4.64 + k_2 \pi_i)$ . To make the range of  $Z$  similar across different mean structures,  $k_1$  takes values of 3 and 6, and  $k_2$  takes values of 26 and 52, when the sampling rate is

10% and 5%, respectively. Figure 1 plots the two populations. We then generated the binary outcome variable  $Y_1$ , where  $Y_1$  is equal to one if  $Z$  is less than or equal to its superpopulation 10<sup>th</sup> percentile, otherwise  $Y_1$  is equal to zero. Similarly, we generated the binary outcomes  $Y_2$  and  $Y_3$  by using the superpopulation 50<sup>th</sup> and 90<sup>th</sup> percentiles of  $Z$  as cut-off values. The target of inference here is the population proportion with  $Y$  equal to one.

In each simulation replicate, a finite population was generated before a sample was drawn, and the true finite population proportion with  $Y$  equal to one was calculated and denoted as  $p$ . A pps sample was then drawn systematically from a randomly ordered list of the finite population. For each population and sample size combination, 1,000 replicates were obtained and the six estimators were compared in terms of empirical bias, root mean squared error (RMSE), and the non-coverage rate of the 95% confidence /credible interval. Simulation results are presented in Tables 1 through 3. Let  $\hat{p}_i$  be an estimate of  $p_i$  based on the  $i^{\text{th}}$  pps sample, the empirical bias and RMSE are defined as follow,

$$\text{Bias} = \frac{1}{1,000} \sum_{i=1}^{1,000} (\hat{p}_i - p_i),$$

$$\text{RMSE} = \sqrt{\frac{1}{1,000} \sum_{i=1}^{1,000} (\hat{p}_i - p_i)^2}.$$

## 5.2 Simulation results

Figure 2 shows the posterior means of  $\Pr(Y_i = 1 | \pi_i)$  and 95% credible intervals based on the Bayesian probit linear  $p$ -spline model for a random pps sample from the EXP case. The upper left plot is the scatter plot of the continuous variable  $Z$  in a pps sample, with three

horizontal parallel lines superimposed, representing the superpopulation 10<sup>th</sup>, 50<sup>th</sup>, and 90<sup>th</sup> percentiles, respectively. In the upper right plot, the binary variable  $Y$ , defined as 1 if  $Z$  is less than or equal to the superpopulation 10<sup>th</sup> percentile, are plotted with black circles, and the superpopulation  $\Pr(Y_i = 1 | \pi_i)$  are plotted with a solid black curve. The solid grey curve and two dashed grey curves are the posterior means of  $\Pr(Y_i = 1 | \pi_i)$  and 95% credible intervals based on the Bayesian probit linear  $p$ -spline regression model. The other two plots are similar to the upper right plot, but with superpopulation 50<sup>th</sup> and 90<sup>th</sup> percentiles as cut-off values in defining  $Y$ . These plots show that the true probabilities of  $Y = 1$  fall within the 95% credible intervals, and are close to the posterior means of  $\Pr(Y_i = 1 | \pi_i)$ . We conclude that the Bayesian probit  $p$ -spline regression model fits well for the binary outcomes in the nonlinear case.

Table 1 shows the empirical bias ( $\times 10^3$ ) for the six estimators in the two populations generated via LINUP and EXP. Overall the design-based estimators (a, d, and f) are less biased than the model-based estimators (b, c, and e). In the LINUP case, the linear probit regression model is correctly specified, so that the empirical bias of the PR estimators are similar to the empirical bias of the BPSP estimator; while in the EXP case, a nonlinear probit regression is needed to fit the data, and thus the PR estimator is more biased than the BPSP estimator when the true population proportions are 0.1 and 0.5. However, the LR estimator has similar to the BPSP estimator empirical bias because of the IBC property. Compared to the model-based PR and BPSP estimators, the PR\_GR and BPSP\_GR estimator reduce the bias by adding the bias calibration term. Moreover, no matter which assisting models were used, both GR estimators achieve similar empirical bias.

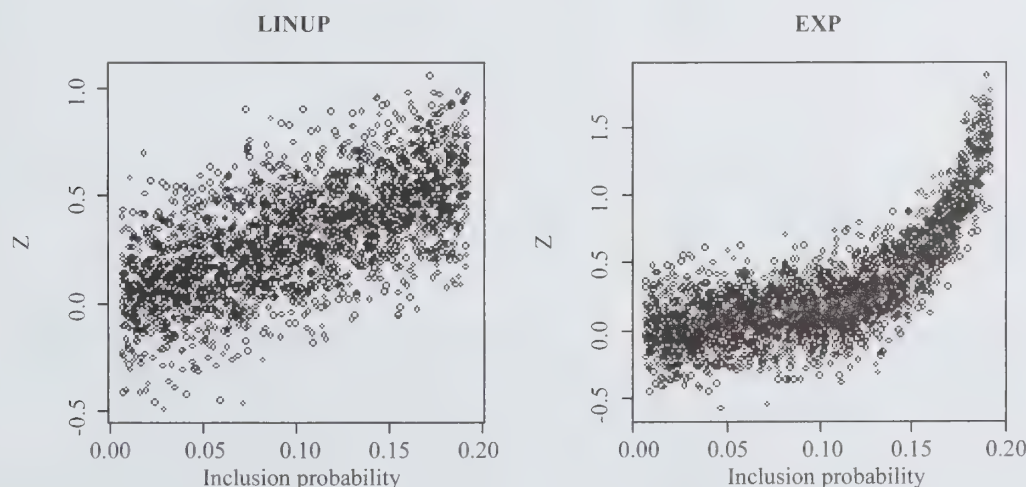
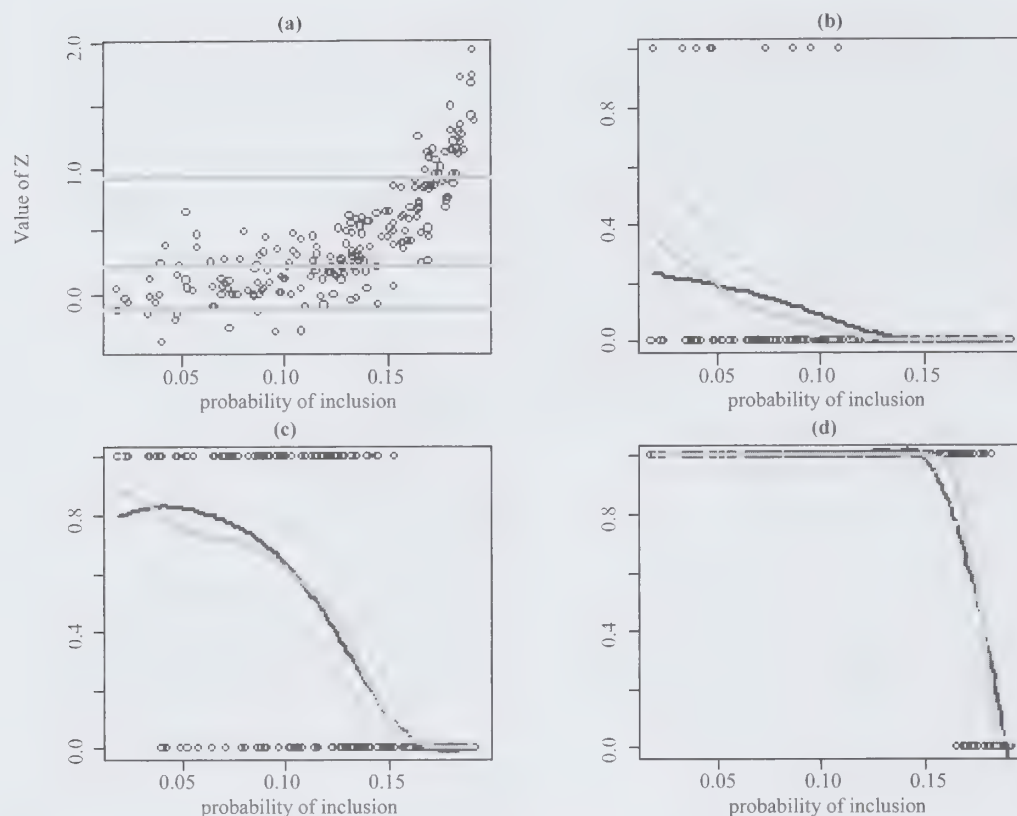


Figure 1 Two simulated artificial populations ( $N = 2,000$ )





**Figure 2** A random pps sample from the EXP case ( $n = 200$ ,  $N = 2,000$ ): (a) scatter plot of  $Z$ ; the three grey lines are the superpopulation 10<sup>th</sup>, 50<sup>th</sup>, and 90<sup>th</sup> percentiles, respectively. (b) black circles are observed units of binary survey variable  $Y$  in the sample, defined as  $Y = I(Z \leq 10^{\text{th}} \text{ percentile})$ ; the grey solid and dashed curves are posterior means of  $\Pr(Y_i = 1 | \pi_i)$  and 95% credible intervals, respectively, simulated based on a probit  $p$ -spline model on  $\pi$ ; and the black curve is the superpopulation  $\Pr(Y_i = 1 | \pi_i)$ . (c) similar to (b), but with  $Y = I(Z \leq 50^{\text{th}} \text{ percentile})$ . (d) similar to (b), but with  $Y = I(Z \leq 90^{\text{th}} \text{ percentile})$

**Table 1**

Empirical bias  $\times 1,000$  of six estimators (Minimum absolute bias within a row is in italic print)

Population	$n$	True prop.	HK	LR	PR	PR_GR	BPSP	BPSP_GR
LINUP	100	0.10	<i>-0.01</i>	13.0	10.3	1.6	8.0	1.2
		0.50	-4.0	-2.9	-4.3	-3.0	-5.2	-3.3
		0.90	-0.4	0.3	-2.5	0.3	-2.9	<i>0.08</i>
	200	0.10	2.5	7.9	5.8	1.5	5.1	<i>1.4</i>
		0.50	3.3	-0.1	-1.3	<i>-0.06</i>	-1.7	-0.2
EXP	100	0.90	1.6	0.4	-1.0	<i>0.3</i>	-1.2	<i>0.3</i>
	200	0.10	<i>1.2</i>	18.1	25.8	4.7	17.0	3.9
		0.50	-4.0	-3.5	12.5	-1.6	<i>-1.4</i>	-3.4
		0.90	-1.3	-0.2	-1.0	<i>-0.1</i>	-1.0	-0.2
	200	0.10	3.1	11.0	22.1	3.5	13.4	2.7
		0.50	3.8	-0.6	14.0	0.4	<i>0.01</i>	-0.7
		0.90	2.3	0.1	-0.7	0.1	-0.7	<i>0.02</i>

Table 2 shows the empirical root mean squared error ( $\times 10^3$ ) for the six estimators. The BPSP estimator has much smaller empirical root mean squared error than the HK estimator, except when  $p$  is 0.1 in the EXP case. Overall the PR estimator performs similarly to the BPSP estimator. To protect against model misspecification, the GR estimators lose some efficiency compared to their corresponding

model-based predictive estimators. The PR\_GR estimator has similar to the BPSP\_GR estimator RMSE, but both of the two GR estimators have smaller RMSE compared to the HK estimator by using assisting models.

Table 3 shows the noncoverage probability ( $\times 10^2$ ) of 95% confidence/credible intervals, the probability that the true finite population proportion is outside the 95% CI of the

estimators. To calculate the variances of estimators, we use the Yates-Grundy variance estimator as defined in equation (2) for the HK estimator; use jackknife resampling method defined by equation (10) for the LR estimator; and use both the linearization (V1) method defined by equation (9) and the jackknife resampling (V2) method for the PR\_GR and BPSP\_GR estimators. Overall, the confidence coverage of credible interval for the BPSP estimator is closer to the nominal level than the other five estimators, especially when the population proportion  $p$  is close to zero or one or when few observations are selected into sample in the tails. Specifically, the BPSP estimator achieves significant improvement in coverage when  $p$  is close to zero in both the LINUP and EXP cases, since little data are included in the sample from the lower tail of the two populations. Note that the improved coverage of the BPSP estimator is achieved with intervals that are narrower on average than those of the HK, LR, PR\_GR, and BPSP\_GR estimators. Similar to the empirical bias and RMSE, the BPSP\_GR does not improve the coverage in comparison to the PR\_GR estimator by using a flexible assisting model.

The choice of prior and hyperprior distributions in mixed models can have a big effect on inferences. We used a prior distribution  $N(0,10^6)$  for the fixed effects parameters,  $\beta_i$ . In our simulations, we report results based on a proper inverse-gamma prior distribution for  $\tau^2$ , namely  $\tau^2 \propto IG(0.1,0.1)$ . To assess sensitivity to the choice of prior distributions, we also computed results using  $\tau^2 \propto IG(0.01,0.01)$  and  $\tau^2 \propto IG(0.001,0.001)$ , as well as an improper uniform prior distribution on  $\tau$  (Gelman 2006). These different priors had little impact on posterior inference of the proportion of interest.

6. Example of tax auditing

We now compare the BPSP estimator with alternative methods on a real population involving income tax auditing data (Compumine 2007). The data set consists of 3,119 Swedish income tax returns for persons who during the year

sold mutual funds managed in a foreign country. The outcome of interest  $Y$  is whether the income tax return is incorrect (coded as 1 for incorrect, and 0 for correct), and it is measured for all observations in this data set. We treated the 3,119 income tax returns as a finite population here, so that the true population proportion of incorrect income tax returns is 0.517. Since the amount of the realized positive profit is an important feature for determining the amount the tax payer has hidden from taxation for his return of income from the sale of a foreign fund, it was chosen as the size variable used in drawing pps sampling. When the primary measure of interest is the total amount the tax payer has hidden from taxation, it is reasonable to assign a value of 1 Swedish Krona to negative profits, the minimum amount of the positive profits, where negative values are not allowed in the size variable.

One thousand repeated systematic pps samples of size 300 and 600 were drawn without replacement from randomly ordered population lists. The returns with largest profits were included with certainty into the samples of size 300 and 600: there were 78 and 241 such returns respectively. Figure 3 shows that the probability of inclusion has a right-skewed distribution for the population even after excluding the observations with inclusion probability of 1.

We applied the same six estimators as in the simulation study with 30 knots on the pps samples, and compared their performances in terms of empirical bias, RMSE, and average width and noncoverage rate of the 95% confidence/credible interval. For the BPSP estimator, a fixed number of 30 knots are placed at evenly spaced sample percentiles of the inclusion probabilities. For the GR estimators, neither the linearization nor the jackknife variance estimator has predominantly better performance than the other, we present the inference based on the linearization variance estimator for simple calculation. We report the GR estimators based on both equations (6) and (7). The results are displayed in Table 4.

Table 2  
Empirical RMSE  $\times$  1,000 of six estimators (Minimum RMSE within a row is in italic print)

Population	<i>n</i>	True prop.	HK	LR	PR	PR_GR	BPSP	BPSP_GR
LINUP	100	0.10	55.1	57.1	46.3	51.3	47.2	51.7
		0.50	65.2	50.8	47.1	49.7	47.7	50.0
		0.90	26.3	22.6	23.3	22.7	23.5	22.9
	200	0.10	39.3	40.9	31.8	36.1	32.0	36.2
		0.50	45.7	35.9	32.8	34.3	32.8	34.6
		0.90	17.8	15.4	15.5	15.4	15.5	15.3
EXP	100	0.10	51.2	60.1	54.4	51.6	51.8	52.4
		0.50	66.1	56.0	43.0	53.2	47.0	51.7
		0.90	24.2	12.4	12.3	12.4	12.3	12.3
	200	0.10	35.9	42.4	39.6	35.6	36.0	36.2
		0.50	45.1	38.9	31.3	36.1	32.1	35.1
		0.90	15.8	8.0	8.1	8.0	8.0	8.0



**Table 3**  
Noncoverage rate of 95% CI  $\times$  100 of six estimators (noncoverage rate within a row closest to 5 is in italic print)

Population	<i>n</i>	True prop.	HK	LR	PR	PR_GR		BPSP	BPSP_GR	
						V1	V2		V1	V2
LINUP	100	0.10	16.2	18.0	8.4	20.9	16.1	9.0	18.4	14.2
		0.50	7.5	9.4	5.0	7.2	7.6	4.4	7.3	7.1
		0.90	7.4	11.4	5.7	8.0	9.4	5.4	8.4	7.1
	200	0.10	10.8	12.6	6.4	13.9	10.9	6.2	12.6	9.4
		0.50	5.5	8.3	5.5	6.2	5.9	5.1	6.0	5.5
		0.90	6.0	8.4	4.4	6.1	4.4	4.7	6.3	5.5
EXP	100	0.10	15.0	18.1	10.5	19.4	14.8	9.2	18.4	14.4
		0.50	7.4	13.5	12.2	9.0	11.4	8.9	10.2	8.4
		0.90	6.1	10.5	7.9	9.9	7.6	7.0	9.8	7.2
	200	0.10	10.8	13.3	9.9	12.5	11.7	7.5	12.4	9.4
		0.50	6.0	11.5	14.3	7.2	8.5	6.2	7.5	6.9
		0.90	5.5	8.8	5.5	6.8	4.6	5.5	6.6	3.7

\* V1: variance estimator using linearization; V2: jackknife variance estimator.

Table 4 shows that the BPSP estimator has slightly increased bias but smaller RMSE, shorter average width and closer to the nominal level credible interval than the design-based estimators (a), (d), and (f). Results not shown here indicate that the BPSP estimator with a uniform prior distribution has slightly better performance than that with inverse-gamma prior distribution with respect to empirical bias, RMSE, and coverage rate, because there are more fluctuations in the data and the uniform prior allows the fitted function to have more flexibility. The BPSP\_GR estimator is less biased, but achieves less efficiency and worse coverage rate than the BPSP estimator. The predictive estimator using the probit linear regression model as prediction model performs poorly here since the model is misspecified, but its GR estimator does reduce bias and RMSE and improve coverage rate. The BPSP\_GR estimator based on equation (6) performs very poorly in terms of RMSE compared to the estimator in equation (7), because a situation similar to that in Basu's (1971) circus elephant example occurs, where one or more observations having very low inclusion probabilities are selected into the sample and hence receive large weights. However, the PR\_GR estimator in equation (6) performs as well as that in equation (7) with predictions obtained from the weighted maximum likelihood estimates, where inclusion probability is used as a covariate as well as the sample weights. Overall, the GR estimator in equation (7) is more desirable than that in equation (6). As the sample size increases from 300 to 600, the noncoverage probability of the 95% credible interval of the BPSP estimator approaches the nominal level of 5% quickly from 14% to 5%, but the coverages are consistently below the nominal level for the other estimators.

Compared to the linear model-based predictive estimators, the BPSP estimator is robust not only to model misspecification, but also to the influential observations in the sample. To demonstrate the robustness to the influential observations, we compare the changes in the model fitting

using probit  $p$ -spline models, linear probit model, and quadratic probit model based on the pps sample only in Figure 4, and based on the pps sample as well as the observations with inclusion probabilities of 1 in Figure 5. In each figure, the population is stratified by the 100 quantiles of the probabilities of inclusion, and the true probabilities of  $Y = 1$  are calculated and plotted with a black dot for each stratum. The grey curves are the posterior means of  $\Pr(Y_i = 1 | \pi_i)$  from 10 random pps samples using 3,000-iterate Gibbs sampler and linear spline in the left plot, using linear probit regression in the middle plot, and using quadratic probit regression in the right plot. Figure 4 shows that the probit  $p$ -spline regression model is more flexible in catching the pattern among the observations than the parametric models. From Figure 4 to Figure 5, the posterior means of  $\Pr(Y_i = 1 | \pi_i)$  do not change except for those with very large inclusion probabilities using the  $p$ -spline model. However, the posterior means curves change dramatically using the quadratic probit regression. These comparisons indicate that probit  $p$ -spline regression model is less likely affected by influential observations, and hence is a good choice of prediction model in the model-based inference.

## 7. Discussion

Bayesian inferences based on the  $p$ -spline model outperform the HK estimator, the GR estimators, and linear model-based prediction estimators in our simulations. The BPSP estimators are more efficient than the HK and GR estimators, and despite slightly higher empirical bias, their 95% credible intervals provide better confidence coverage and shorter average interval width, especially when the population proportion is closer to zero or one and few data are selected into the sample in the tails. This suggests the importance of current research in estimating finite population prevalence of rare events.

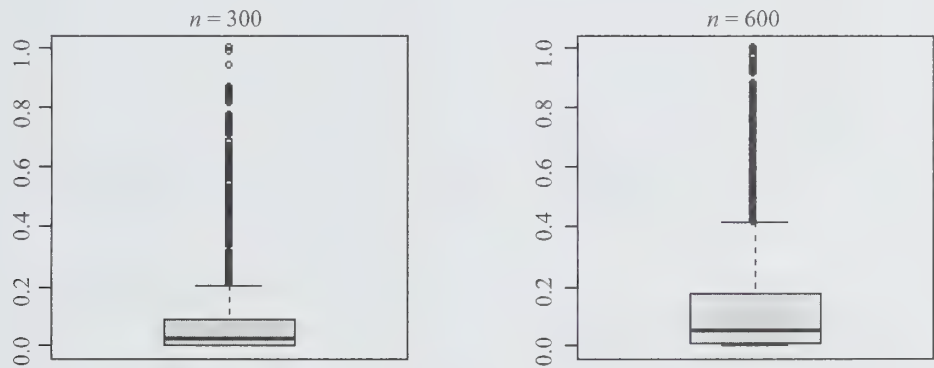
The BPSP estimator is a natural extension of the regular linear regression model-based estimators of finite population proportions. Compared to linear model-based predictive estimators, the BPSP estimator achieves robustness to model misspecification and influential observations in the sample by using a flexible  $p$ -spline model, without much

loss of efficiency for the sample sizes considered. Therefore, the BPSP estimator is easy to understand while requires complex computation. However, with the availability of WinBUGS, the Bayesian statistical software, the BPSP estimator can be easily implemented by survey practitioners.

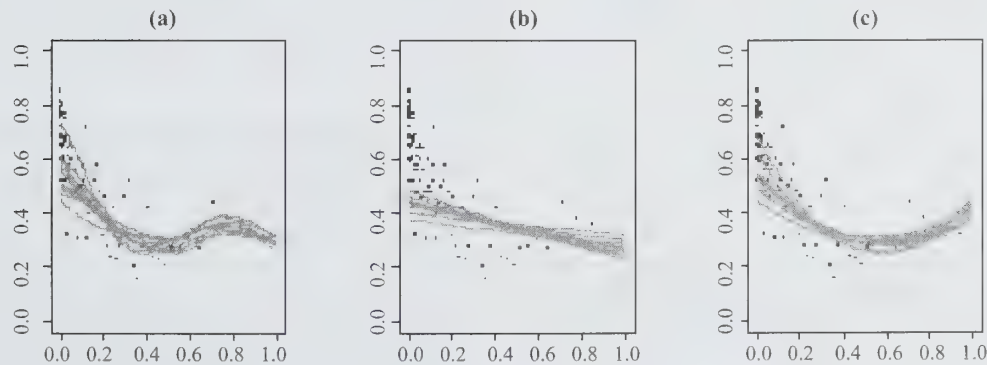
**Table 4**  
Comparison of various estimators for empirical bias, root mean squared error, and average width and noncoverage rate of 95% CI, in the tax return example

Methods	bias*100		RMSE*100		average width*100		noncoverage*100	
	300	600	300	600	300	600	300	600
HK	-2.4	-1.8	12.4	10.2	36	29	14.1	10.2
LR	6.7	5.5	11.9	9.2	27	21	43.5	45.6
PR	-11.6	-10.1	12.4	10.6	18	14	69.8	83.4
PR_GR1	-1.2	-0.4	11.5	8.7	31	25	22.4	16.8
PR_GR2	-1.2	-0.3	11.5	8.8	33	26	16.1	11.4
BPSP	-6.8	-2.7	9.3	5.2	27	19	14.2	5.0
BPSP_GR1	-3.0	-0.5	102.6	56.9	77	57	14.4	9.2
BPSP_GR2	-0.7	0.2	12.0	10.1	34	26	15.9	12.8

\* GR\_1: GR estimators using equation (6);  
GR\_2: GR estimators using equation (7).

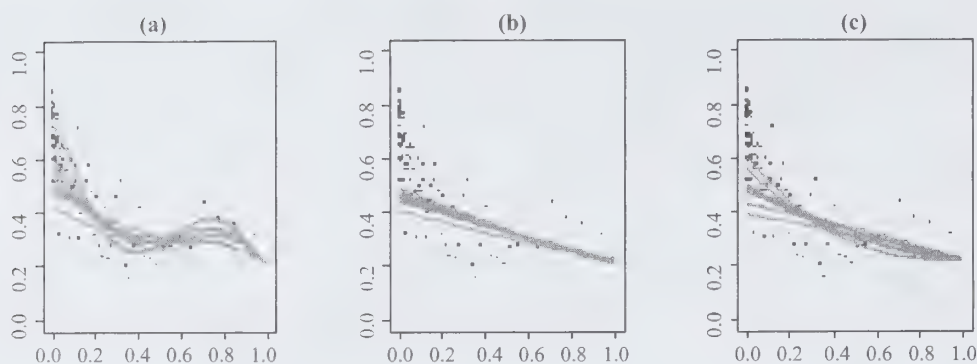


**Figure 3** Box plots of the probabilities of inclusion for two sample sizes in the tax auditing example



**Figure 4** Predictions based on pps samples only in the tax auditing example, X-axis: inclusion probabilities  $\pi$ , Y-axis:  $P(Y = 1|\pi)$ ; black dots are the true  $P(Y = 1|\pi)$  within each percentile of  $\pi$ ; grey curves are ten realizations of the posterior means of  $P(Y = 1|\pi)$ . The prediction models are (a) probit linear  $p$ -spline regression, (b) linear probit regression, (c) quadratic probit regression





**Figure 5** Predictions based on the combined data of pps samples and the observations sampled with certainty in the tax auditing example, X-axis: inclusion probabilities  $\pi$ , Y-axis:  $P(Y=1|\pi)$ ; black dots are the true  $P(Y=1|\pi)$  within each percentile of  $\pi$ ; grey curves are ten realizations of the posterior mean of  $P(Y=1|\pi)$ . The prediction models are (a) probit linear  $p$ -spline regression, (b) linear probit regression, (c) quadratic probit regression

The BPSP estimators are not sensitive to two choices of prior distributions of  $\tau^2$  considered here, though it appears from the tax auditing example that the uniform prior yields slightly smaller bias and RMSE, shorter 95% credible intervals, and better coverage when a nonlinear prediction model is needed. The tax auditing example also shows that in the GR estimator, an estimated population size using the sum of inverse inclusion probabilities is more desirable than the true population size when one or more observations with very low inclusion probability are included in the sample, since the GR estimator with denominator  $N$  has high variance and low efficiency in this case.

The design-based estimators and their 95% confidence intervals can provide valid inferences for population proportions when the sample is large. However, these asymptotic properties do not appear to hold when the sample size is moderate or small. The BPSP approach can provide more valid inferences for small samples, especially when the true population proportion to be estimated is close to 0 or 1, although confidence coverage appears to be less than nominal when the sample size gets small, and lack of parsimony of the model is an issue. When estimating proportions away from tails, the BPSP estimator leads to slightly smaller RMSE and closer to the nominal level confidence coverage than the HK and GR estimators, but the improvement is not so significant as in the tails. In this scenario, to avoid the complex computation of the BPSP estimator, the PR\_GR estimator based on equation (7) is an alternative to the survey practitioners.

The choice of variance estimator is problematic for some unequal probability designs for the design-based estimators, but the Bayesian  $p$ -spline prediction approach provides a simulation approximation of the full posterior distribution of

the population proportion. Extra work is not needed to estimate the variance or 95% credible interval for the BPSP estimator, as it can be obtained simultaneously with the point estimators. In Zheng and Little (2005), three variance estimators of the  $p$ -spline model-based estimator for finite population total in a pps sample were compared, including the model-based empirical Bayes variance estimator, the jackknife variance estimate, and the balanced repeated replication (BRR) variance estimate. The simulation studies showed that the jackknife method worked well, whereas the BRR method tended to yield conservative standard errors and the model-based empirical Bayes estimator was vulnerable to misspecification of the variance structure. In the present work, the  $1 - \alpha$  level credible interval for the BPSP estimator of population proportion is constructed by splitting  $\alpha$  equally between the upper and lower endpoints of the posterior distribution of  $p$ . This pure Bayesian approach based on draws from the posterior distributions seems to work well in our setting and avoids the heavy computation associated with the jackknife and BRR method.

The BPSP estimator we propose here can be extended to include additional auxiliary covariates by adding linear terms for these variables. For domain estimation, an interaction term between the spline of inclusion probabilities and the domain indicator should also be modeled. Both the additive effects of auxiliary variables and the interaction between the domain indicator and inclusion probabilities can be represented in a mixed model (Ruppert *et al.* 2003, page 231) and estimated using Gibbs sampling or WinBUGS (Crainiceanu *et al.* 2005). The BPSP estimator for finite population proportions can also be extended to a more general case of a polychotomous response. The Gibbs

sampling approach for the binary case can be generalized to the case of ordered categories, and can be applied to the unordered categories with a latent multinomial distribution (Albert and Chib 1993). Another extension for the BPSP estimator is in the small area estimation, by combining small area random effects with the smooth spline on the inclusion probabilities (Opsomer, Claeskens, Ranalli, Kauermann and Breidt 2008). This extension will be the focus of future research.

Finally, one reviewer questioned whether the proposed approach can be applied in a multipurpose survey with many outcomes, since the modeling procedure does not provide a single set of weights and needs to be repeated for all variables of interest. It is true that our methods are more computationally intensive than existing approaches, but the BPSP method can be easily implemented with a Gibbs sampling algorithm or using WinBUGS, so computing is not a major obstacle. We point out that the simulations in the paper involved repeating the iterative Gibbs analysis 6,000 times, so an equivalent level of computation on a single survey of comparable size would allow the implementation of the BPSP method for 6,000 outcomes! These were done on a garden-variety laptop PC. While we do not advocate automatic use of any analytical method, design or model-based, our point is that computational complexity is no longer a major obstacle to applying these methods. We suggest that the statistical properties of a method are more important than computing time, given modern day computing resources.

### Acknowledgements

This work is supported in part by The Dow Chemical Company through an unrestricted grant to the University of Michigan Dioxin Exposure Study. The authors thank the referees and an associate editor for their helpful comments on the original version of this paper.

### Appendix

#### Algorithm of Gibbs sampling

Model (3) can also be written in the matrix form,

$$\Phi^{-1}(E(y_i | \beta, b, X, Z)) = (X\beta + Zb), i = 1, \dots, n$$

$$\beta = (\beta_0, \beta_1, \dots, \beta_p)^T, b = (b_1, \dots, b_m)^T \sim N_m(0, \tau^2 I_m)$$

$$X = \begin{pmatrix} 1 & \pi_1 & \dots & \pi_1^p \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 1 & \pi_n & \dots & \pi_n^p \end{pmatrix}, Z = \begin{pmatrix} (\pi_1 - k_1)_+^p & \dots & (\pi_1 - k_m)_+^p \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ (\pi_n - k_1)_+^p & \dots & (\pi_n - k_m)_+^p \end{pmatrix}.$$

The algorithm of Gibbs sampling for estimating the parameters in Model (3) is as follows:

- The probit regression model for the binary outcome  $y = [y_1, \dots, y_n]^T$  corresponds to a normal regression model for a latent continuous data  $y^* = [y_1^*, \dots, y_n^*]^T$ , which has a truncated multivariate normal distribution with mean  $(X\beta + Zb)$  and identity covariance matrix (Albert and Chib 1993), and  $y_i$  is the indicator that  $y_i^* > 0$ . With some initial values of  $(\beta, b)$ , values of the latent continuous data  $y_i^*$  can be simulated.
- Specifying a proper flat normal prior distribution  $N(0, 10^6)$  on  $\beta$  and an inverse gamma distribution  $IG(0.1, 0.1)$  on  $\tau^2$ , the posterior distribution of  $(\beta, b, \tau^2)$  given the simulated latent continuous data  $y^*$  is

$$(\beta, b) | \tau^2, y^* \sim \text{MVN}_{m+p+1}((C^T C + D/\tau^2)^{-1} C^T y^*, (C^T C + D/\tau^2)^{-1})$$

$$\tau^2 | \beta, b \sim \text{IG}(0.1 + m/2, 0.1 + \|b\|^2/2), \quad (11)$$

where  $C = [X, Z]$  and  $D$  is a diagonal matrix with  $p+1$  values of  $10^{-6}$  followed by  $m$  ones on the diagonal. Gelman (2006) recommended a uniform prior distribution on  $\tau$ , which results in the posterior distribution for  $\tau^2$  as

$$\tau^2 | \beta, b \sim \text{IG}((m-1)/2, \|b\|^2/2) \quad (12)$$

- At iteration  $t$ , draws of  $(\beta^{(t)}, b^{(t)}, \tau^{2(t)})$  from the posterior distribution in equation (11) or (12) are used to generate new latent data  $\hat{y}^{*(t)}$  conditional on observed binary variable  $y$  for the sample, and to obtain the posterior predicted values  $\hat{y}^{(t)}$  for non-sample units. We then can obtain draws from the posterior distribution of the finite population proportion at iteration  $t$  as

$$\hat{p}_{\text{PR}}^{(t)} = N^{-1} \left( \sum_{i \in S} y_i + \sum_{j \notin S} \hat{y}_j^{(t)} \right)$$

### References

- Albert, J.H., and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of American Statistical Association*, 88, 669-679.
- Basu, D. (1971). An essay on the logical foundations of survey sampling. Part 1, in *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.



- Compumine (2007). Re: analysis – Tax audit data mining. Feb. 2007. <http://www.compumine.com/web/public/newsletter/20071/tax-audit-data-mining>.
- Crainiceanu, C.M., Ruppert, D. and Wand, M. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, 14, 2005, 14.
- Duchesne, P. (2003). Estimation of a proportion with survey data. *Journal of Statistics Education*, 11, 3.
- Eilers, P.H.C., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11, 89-121.
- Firth, D., and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, 60, 3-21.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3, 515-533.
- Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- Horvitz, D.G., and Thompson, M.E. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.
- Isaki, C.T., and Fuller, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2, 813-830.
- Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-673.
- Lehtonen, R., and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24, 51-55.
- Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.
- Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 24, 69-77.
- Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B*, 70, 265-286.
- Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.
- Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Shao, J., and Wu, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.
- Smith, T.M.F. (1976). The foundations of survey sampling: A review (with discussion). *Journal of the Royal Statistical Society, Series A*, 139, 183-204.
- Smith, T.M.F. (1994). Sample surveys 1975-1990: An age of reconciliation? (with discussion). *International Statistical Review*, 62, 5-34.
- Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B*, 15, 235-261.
- Zheng, H., and Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.
- Zheng, H., and Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

# The effect of nonresponse adjustments on variance estimation

David Haziza, Katherine Jenny Thompson and Wesley Yung<sup>1</sup>

## Abstract

Many surveys employ weight adjustment procedures to reduce nonresponse bias. These adjustments make use of available auxiliary data. This paper addresses the issue of jackknife variance estimation for estimators that have been adjusted for nonresponse. Using the reverse approach for variance estimation proposed by Fay (1991) and Shao and Steel (1999), we study the effect of not re-calculating the nonresponse weight adjustment within each jackknife replicate. We show that the resulting 'shortcut' jackknife variance estimator tends to overestimate the true variance of point estimators in the case of several weight adjustment procedures used in practice. These theoretical results are confirmed through a simulation study where we compare the shortcut jackknife variance estimator with the full jackknife variance estimator obtained by re-calculating the nonresponse weight adjustment within each jackknife replicate.

**Key Words:** Calibration; Nonresponse adjustment; Unit nonresponse; Jackknife variance estimator; Linearization variance estimator.

## 1. Introduction

Unit nonresponse, which occurs when, for a sample unit, all the survey variables are missing or when not enough usable information is available, is unavoidable in surveys. To address this, the nonrespondents are deleted from the data file and the survey weights of the respondents are adjusted to compensate for the deletions. The primary objective of a weight adjustment procedure is to reduce the nonresponse bias, which is introduced when respondents and nonrespondents are different with respect to the survey variables. Key to achieving an efficient bias reduction is the use of powerful auxiliary information available for both respondents and nonrespondents.

In this paper, we consider jackknife variance estimation in the presence of unit nonresponse. This variance estimation method is widely used in practice because of its theoretical properties and computational ease. In contrast to Taylor linearization procedures, the jackknife method does not require a separate derivation for each parameter of interest nor the second-order inclusion probabilities that may be difficult to obtain in complex surveys. When using a jackknife variance estimator in the context of nonresponse, there is some question of whether or not the nonresponse adjustment needs to be replicated (e.g., Valliant 2004). In this paper, we consider two jackknife variance estimators: (i) a *full* jackknife variance estimator which recalculates the nonresponse adjustment factor within each jackknife replicate and (ii) a *shortcut* jackknife variance estimator, which does not. The shortcut jackknife variance estimator is convenient in practice but its theoretical properties were not, to our knowledge, fully studied in the literature. Production reasons tend to drive the usage of a shortcut jackknife

variance estimator, since the full jackknife variance estimator in the context of stratified sampling can be quite time-consuming and computer resource-intensive, especially when a survey utilizes a large number of weighting cells. Some recent studies conducted at the U.S. Census Bureau (Thompson 2005 and Ozcoskun, Thompson and Williams 2005) found negligible differences between variance estimates obtained using a fully replicated weight adjustment procedure and those obtained using a "shortcut" procedure with stratified jackknife, delete-a-group jackknife, and modified half sample variance estimators.

Two types of adjustment procedures are commonly used in practice. The first, called *nonresponse propensity weighting* (NPW), consists of first modeling the response propensities and using the inverse of the estimated propensities as the weighting adjustment. The estimated response propensities are typically obtained by fitting a parametric model (e.g., logistic regression model) or by fitting a nonparametric model; e.g., Da Silva and Opsomer (2006). A special case of NPW, which is very popular in practice, consists of first dividing the respondents and nonrespondents into weighting classes and adjusting the design weights of respondents by the inverse of the response rate within each class. These classes are formed on the basis of auxiliary information recorded for all units in the sample; see, for example, Eltinge and Yansaneh (1997) and Little (1986). The second type of adjustment procedures, called *nonresponse calibration weighting* (NCW) can be seen as an extension of the calibration approach (Deville and Särndal 1992) adapted to the context of unit nonresponse. The reader is referred to Särndal and Lundström (2005), Kott (2006) and Brick and Montaquila (2008) for a comprehensive overview of NPW and NWC. In some

1. David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, H3C 3J7, Canada. E-mail: David.haziza@umontreal.ca; Katherine Jenny Thompson, U.S. Census Bureau, Washington, DC 20233. E-mail: Katherine.J.Thompson@census.gov; Wesley Yung, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: wesley.yung@statcan.gc.ca.



situations, NPW and NCW lead to the same estimator; for example, the count-adjusted estimator presented below (see expression (1.4)). In this paper, we focus on NCW. The problem of variance estimation in the context of NPW has been recently studied by Kim and Kim (2007).

Consider a finite population  $U$  of size  $N$ . The objective is to estimate the population total  $Y = \sum_{i \in U} y_i$ , of a variable of interest  $y$ . Suppose that a random sample  $s$  of size  $n$  is selected from  $U$  according to a given design  $p(s)$ . In the case of complete data, a basic estimator of  $Y$  is the well-known expansion estimator given by

$$\hat{Y}_\pi = \sum_{i \in s} d_i y_i \quad (1.1)$$

where  $d_i = 1/\pi_i$  denotes the design weight attached to unit  $i$  and  $\pi_i = P(i \in s)$  denotes its first-order probability of inclusion in the sample. In the presence of unit nonresponse, only a subset of  $s$  is observed, and so the computation of  $\hat{Y}_\pi$  in (1.1) is not possible.

To define a nonresponse adjusted estimator of  $Y$ , we assume that a vector of auxiliary variables  $\mathbf{x}$  is available for all the sampled units (respondents and nonrespondents) so that the vector of estimated totals,  $\hat{\mathbf{X}}_\pi = \sum_{i \in s} d_i \mathbf{x}_i$ , is available. We also assume that a vector of instrumental variables  $\mathbf{z}$ , of the same dimension as  $\mathbf{x}$ , is available for the respondents. Let  $r_i$  be a response indicator attached to unit  $i$  such that  $r_i = 1$  if unit  $i$  is a responding unit and  $r_i = 0$ , otherwise. To estimate  $Y$ , we consider calibration estimators of the form

$$\hat{Y}_{\text{CAL}} = \sum_{i \in s} w_i r_i y_i, \quad (1.2)$$

where  $w_i = d_i g_i$  and  $g_i$  is a nonresponse weighting adjustment factor attached to unit  $i$  and given by

$$g_i = 1 + (\hat{\mathbf{X}}_\pi - \hat{\mathbf{X}}_r)' \hat{\mathbf{T}}_r^{-1} \mathbf{z}_i, \quad (1.3)$$

where  $\hat{\mathbf{X}}_r = \sum_{i \in s} d_i r_i \mathbf{x}_i$  and  $\hat{\mathbf{T}}_r = \sum_{i \in s} d_i r_i \mathbf{z}_i \mathbf{x}_i'$ . When  $\mathbf{z}_i = \mathbf{x}_i/v_i$ , where  $v_i$  is a known constant, then the estimator (1.3) is identical to the *InfoS* estimator given in Särndal and Lundström (2005, equation 7.15). The properties of the estimator (1.2) were studied by Deville (2002), Sautory (2003), Särndal and Lundström (2005) and Kott (2006), among others.

In this paper, the properties (e.g., bias and variance) of  $\hat{Y}_{\text{CAL}}$  are studied using the nonresponse model (NM) approach, under which inference is made with respect to the joint distribution induced by the sampling design and the nonresponse mechanism,  $q(\mathbf{r} | \mathbf{I})$ , where  $\mathbf{I} = (I_1, \dots, I_N)'$  is the vector of sample selection indicators such that  $I_i = 1$  if unit  $i$  is selected in the sample and  $I_i = 0$ , otherwise and  $\mathbf{r} = (r_1, \dots, r_N)'$  is the vector of response indicators. Let  $p_i = P(r_i = 1 | \mathbf{I}, I_i = 1)$  be the response probability for

unit  $i$ . We assume that  $p_i > 0$  for all  $i$  and that the units respond independently of one another; that is,  $p_{ij} = P(r_i = 1, r_j = 1 | \mathbf{I}, I_i = 1, I_j = 1, i \neq j) = p_i p_j$ .

The estimator  $\hat{Y}_{\text{CAL}}$  is asymptotically unbiased for the true total  $Y$  if (i)  $p_i^{-1} = 1 + \boldsymbol{\lambda}' \mathbf{z}_i$  for all  $i \in U$ , where  $\boldsymbol{\lambda}$  is a vector of unknown constants or (ii)  $y_i = \mathbf{x}_i' \boldsymbol{\beta}$  for all  $i \in U$ , where  $\boldsymbol{\beta}$  is a vector of constants; see Särndal and Lundström (2005, chapter 9.5). If the condition (i) is satisfied, the point estimator  $\hat{Y}_{\text{CAL}}$  is asymptotically unbiased for  $Y$  regardless of the variable of interest  $y$  being estimated. Also, it follows from (ii) that  $\hat{Y}_{\text{CAL}}$  has a small bias if the residuals  $E_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$ , are small, where  $\boldsymbol{\beta} = (\sum_{i \in U} \mathbf{z}_i \mathbf{x}_i')^{-1} \sum_{i \in U} \mathbf{z}_i y_i$ . Therefore, the bias of the estimator  $\hat{Y}_{\text{CAL}}$  is small if the vector  $\mathbf{x}$  explains the variable of interest  $y$ . In the case of several variables of interest, note that the vector  $\mathbf{x}$  may explain a given variable of interest well but may not be related to all, in which case some estimates could be potentially biased. We assume that  $\hat{Y}_{\text{CAL}}$  is asymptotically unbiased for  $Y$ , so that the bias of the estimators under consideration is not an issue in the remainder of the paper.

We consider three special cases of (1.2) that are of interest in practice (see also Kalton and Flores-Cervantes 2003). First, let  $\boldsymbol{\delta} = (\delta_{i1}, \dots, \delta_{ic}, \dots, \delta_{iC})'$  be a  $C$ -vector of weighting class indicators attached to unit  $i$  such that  $\delta_{ic} = 1$  if unit  $i$  belongs to class  $c$  and  $\delta_{ic} = 0$ , otherwise for  $c = 1, \dots, C$ . If  $\mathbf{x}_i = \mathbf{z}_i = \boldsymbol{\delta}_i$ , the adjustment factor  $g_i$  given by (1.3) reduces to  $g_i = \hat{N}_c / \hat{N}_{rc} \delta_{ic}$ , where  $\hat{N}_c = \sum_{i \in s} d_i \delta_{ic}$  and  $\hat{N}_{rc} = \sum_{i \in s} d_i r_i \delta_{ic}$ . That is, the nonresponse weighting adjustment factor for a weighting cell is calculated as the sample-weighted number of sampled units in the weighting cell divided by the sample-weighted number of responding units in the weighting cell. We refer to this weight adjustment procedure as the *count adjustment* procedure. It follows that the estimator (1.2) reduces to the count adjusted estimator

$$\hat{Y}_{\text{count}} = \sum_{c=1}^C \frac{\hat{N}_c}{\hat{N}_{rc}} \hat{Y}_{rc}, \quad (1.4)$$

where

$$\hat{Y}_{rc} = \sum_{i \in s} d_i r_i \delta_{ic} y_i.$$

The second special case of (1.2) assumes that a continuous variable  $x$  is available for all the sampled units. Let  $\mathbf{x}_i = (\delta_{i1} x_i, \dots, \delta_{ic} x_i, \dots, \delta_{iC} x_i)'$  and  $\mathbf{z}_i = \boldsymbol{\delta}_i$ . In this case, the adjustment factor  $g_i$  given by (1.3) reduces to  $g_i = \hat{X}_c / \hat{X}_{rc} \delta_{ic}$  if unit  $i$  belongs to class  $c$ , where  $\hat{X}_c = \sum_{i \in s} d_i \delta_{ic} x_i$  and  $\hat{X}_{rc} = \sum_{i \in s} d_i r_i \delta_{ic} x_i$ . Here, the nonresponse weighting adjustment factor for a weighting class  $c$  is the sum of the sample-weighted auxiliary data for units in the weighting cell divided by the sum of the

sample-weighted auxiliary data for all responding units in the weighting cell. We refer to this weight adjustment procedure as the *ratio adjustment* procedure. The estimator (1.2) reduces to the ratio adjusted estimator

$$\hat{Y}_{\text{ratio}} = \sum_{c=1}^C \frac{\hat{X}_c}{\bar{X}_{rc}} \hat{Y}_{rc}. \quad (1.5)$$

Note that the count adjusted estimator (1.4) is a special case of the ratio adjusted estimator when  $x_i = 1$  for all the sampled population units.

Finally, if  $\mathbf{x}_i = \mathbf{z}_i = (\delta_{i1}, \dots, \delta_{ic}, \dots, \delta_{iC}, \delta_{i1}x_i, \dots, \delta_{ic}x_i, \dots, \delta_{iC}x_i)'$ , we obtain another special case of (1.2). In this case, the adjustment factor  $g_i$  given by (1.3) reduces to

$$g_i = \hat{N}_c \left[ 1 + (\bar{x}_c - \bar{x}_{rc}) \frac{(x_i - \bar{x}_{rc})}{\sum_{i \in s} r_i \delta_{ic} (x_i - \bar{x}_{rc})^2} \right],$$

if unit  $i$  belongs to class  $c$ , where  $\bar{x}_c = \hat{X}_c / \hat{N}_c$  and  $\bar{x}_{rc} = \hat{X}_{rc} / \hat{N}_{rc}$ . We refer to this weight adjustment procedure as the *simple linear regression adjustment* procedure. The estimator (1.2) reduces to the simple linear regression adjusted estimator

$$\hat{Y}_{\text{slreg}} = \sum_{c=1}^C \hat{N}_c [\hat{Y}_{rc} + (\bar{x}_c - \bar{x}_{rc}) \hat{B}_{rc}], \quad (1.6)$$

where

$$\hat{B}_{rc} = \frac{\sum_{i \in s} d_i r_i \delta_{ic} (x_i - \bar{x}_{rc}) (y_i - \bar{y}_{rc})}{\sum_{i \in s} d_i r_i \delta_{ic} (x_i - \bar{x}_{rc})^2}.$$

The estimators (1.4)–(1.6) use some form of weighting adjustment within classes. All of them are asymptotically unbiased for  $Y$  if the units have equal response probabilities within classes (*i.e.*, a uniform nonresponse mechanism within classes). This condition is a special case of condition (i) discussed above.

In this paper, we show that the shortcut jackknife variance estimator that treats the adjustment factors as fixed, tends to overestimate the true variance of  $\hat{Y}_{\text{CAL}}$ , at least in some simple cases. We build on earlier research by Thompson and Yung (2006) who derived expressions of the linearization version for both the full and shortcut jackknife variance estimators and evaluated these expressions empirically using data from the Annual Capital Expenditures Survey (ACES), conducted at the U.S. Census Bureau. In the context of NPW, it is interesting to note that Kim and Kim (2007) showed that treating the estimated response probabilities as fixed leads to an overestimation of the true variance when the sampling weights are not used in estimating these probabilities. Beaumont (2005) obtained similar results in the context of imputation when the response probabilities are estimated using a logistic regression model.

In Section 2, we discuss the full and shortcut jackknife variance estimators and show that the shortcut estimator is asymptotically biased. The severity of this bias is evaluated for two commonly used sample designs in Section 3. Section 4 presents the results of a simulation study comparing the full and shortcut jackknife variance estimators. We conclude in Section 5 with some general observations.

## 2. Jackknife variance estimation

Traditionally, variance estimation in the context of nonresponse has been performed using the two-phase framework, which consists of viewing nonresponse as a second-phase of selection. Instead, we consider the reverse framework that was proposed by Fay (1991) and further developed by Shao and Steel (1999). This framework provides a theoretical basis for studying the properties of jackknife variance estimators and can be described as follows: first, applying the nonresponse mechanism, the population  $U$  is randomly divided into a population of respondents  $U_r$  and a population of nonrespondents  $U_m$ . Then, given  $(U_r, U_m)$ , the random sample  $s$  is selected according to the chosen sampling design. The total variance of  $\hat{Y}_{\text{CAL}}$  can be expressed as

$$V(\hat{Y}_{\text{CAL}}) = E_q V_p(\hat{Y}_{\text{CAL}} | \mathbf{r}) + V_q E_p(\hat{Y}_{\text{CAL}} | \mathbf{r}), \quad (2.1)$$

where  $E_p(\cdot)$  and  $V_p(\cdot)$  denote the expectation and the variance with respect to the sampling design and  $E_q(\cdot)$  and  $V_q(\cdot)$  denote the expectation and variance with respect to the nonresponse mechanism,  $q(\mathbf{r} | \mathbf{I})$ .

In this section, we focus on stratified simple random sampling, which is the design typically used in business surveys. With this sample design, the population  $U$  is partitioned into  $L$  strata  $U_1, \dots, U_L$  of size  $N_1, \dots, N_L$ , respectively. A simple random sample without replacement  $s_h$ , of size  $n_h$ , is selected from stratum  $h$ ,  $h = 1, \dots, L$ . Each within-stratum sample is selected independently, and we assume that  $n_h \geq 2$  for all  $h$ . In this context, the design weight of unit  $i$  in stratum  $h$  is  $d_{hi} = N_h/n_h$ . A full jackknife variance estimator of  $\hat{Y}_{\text{CAL}}$ , under stratified simple random sampling, is obtained as follows:

- (i) remove unit  $(gj)$  from the sample,  $g = 1, \dots, L$ ;  $j = 1, \dots, n_g$ ;
- (ii) adjust the design weights  $d_{hi}$  to obtain the jackknife weights  $d_{hi(gj)}$ , where  $d_{hi(gj)}$  is given by

$$d_{hi(gj)} = \begin{cases} 0 & \text{if } (hi) = (gj) \\ \frac{n_g}{n_g - 1} d_{gi} & \text{if } h = g, i \neq j \\ d_{hi} & \text{otherwise} \end{cases}$$



- (iii) compute the estimator  $\hat{Y}_{\text{CAL}(g,j)}$  in the same way as  $\hat{Y}_{\text{CAL}}$  with the jackknife weights  $d_{hi(g,j)}$  instead of the design weights  $d_{hi}$ ; that is,  $\hat{Y}_{\text{CAL}(g,j)} = \sum_{(hi) \in s} w_{hi(g,j)} r_{hi} y_{hi}$ , where  $w_{hi(g,j)} = d_{hi(g,j)} g_{hi} r_{hi} y_{hi}$  with  $g_{hi(g,j)} = 1 + (\hat{\mathbf{X}}_{\pi(g,j)} - \hat{\mathbf{X}}_{r(g,j)})' \hat{\mathbf{T}}_{r(g,j)}^{-1} \mathbf{z}_{hi}$ ,  $\hat{\mathbf{X}}_{\pi(g,j)} = \sum_{i \in s} d_{hi(g,j)} \mathbf{x}_{hi}$ ,  $\hat{\mathbf{X}}_{r(g,j)} = \sum_{(hi) \in s} d_{hi(g,j)} r_{hi} \mathbf{x}_{hi}$  and  $\hat{\mathbf{T}}_{r(g,j)} = \sum_{(hi) \in s} d_{hi(g,j)} r_{hi} \mathbf{z}_{hi} \mathbf{x}_{hi}'$ .
- (iv) replace the unit deleted in step (i) back into the sample;
- (v) repeat steps (i)-(iv) for all  $(g,j)$  units,  $g = 1, \dots, L$ ;  $j = 1, \dots, n_h$ .

Note that the nonresponse adjustment factors  $g_{hi}$  are recalculated in each replicate. This leads to the full jackknife variance estimator

$$v_{JF} = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j \in s_h} (\hat{Y}_{\text{CAL}(g,j)} - \hat{Y}_{\text{CAL}})^2. \quad (2.2)$$

The variance estimator  $v_{JF}$  is an estimator of the first term on the right hand side of (2.1),  $E_q V_p(\hat{Y}_{\text{CAL}} | \mathbf{r})$ . This term represents the design variance that we would have obtained if the responding units were selected using stratified simple random sampling with replacement, or equivalently, if the stratum sampling fractions,  $(n_h / N_h)$  are negligible. In other words, the full jackknife variance estimator (2.2) is an estimator of the sampling variance conditional on the vector of response indicators  $\mathbf{r}$ . Therefore,  $v_{JF}$  is asymptotically unbiased and consistent for  $E_q V_p(\hat{Y}_{\text{CAL}} | \mathbf{r})$  under stratified simple random sampling with replacement sampling regardless of the validity of the underlying assumptions. Note that since  $v_{JF}$  is an estimator of a sampling variance, it can be readily obtained using software designed for complete-data jackknife variance estimation. In other words, no specialized software is needed. Also, note that the second term on the right hand side of (2.1),  $V_q E_p(\hat{Y}_{\text{CAL}} | \mathbf{r})$ , is not accounted for. Thus, the full jackknife variance estimator does not track the second term in (2.1). However, the contribution of this term to the total variance is negligible if the stratum sampling fractions,  $n_h / N_h$ , are negligible. As a result,  $v_{JF}$  is asymptotically unbiased and consistent for the total variance,  $V(\hat{Y}_{\text{CAL}})$ . That is,  $E_{pq}(v_{JF}) \approx V(\hat{Y}_{\text{CAL}})$ . Since the goal of the research is to compare the full and shortcut jackknife estimators, in the remainder of the paper, we assume that the stratum sampling fractions are negligible and focus on estimates of totals, so that we can omit the estimation of the second term in (2.1). We note that even if the second term is not negligible, our comparisons are valid as both the full jackknife and shortcut estimators would underestimate the total variance by the same term.

A shortcut jackknife variance estimator of  $\hat{Y}_{\text{CAL}}$  is given by

$$v_{JS} = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j \in s_h} (\hat{Y}_{\text{CAL}(g,j)}^* - \hat{Y}_{\text{CAL}})^2, \quad (2.3)$$

where  $\hat{Y}_{\text{CAL}(g,j)}^* = \sum_{(hi) \in s} d_{hi(g,j)} g_{hi} r_{hi} y_{hi}$ . Note that the nonresponse weighting adjustment factors  $g_{hi}$  are not recalculated in each jackknife replicate. In other words, the factors  $g_{hi}$  are treated as constants, which is inappropriate since they depend on the sample and the set of respondents. Therefore, we have  $E_{pq}(v_{JS}) \neq V(\hat{Y}_{\text{CAL}})$ , in general, and the shortcut variance estimator,  $v_{JS}$ , is biased.

To study the magnitude of the bias of  $v_{JS}$ , we consider the difference of the two jackknife variance estimators,  $D = v_{JS} - v_{JF}$ . Since the variance estimator  $v_{JF}$  is an asymptotically unbiased estimator of the term  $V_p(\hat{Y}_{\text{CAL}} | \mathbf{r})$ , it is asymptotically equivalent to a variance estimator obtained using a first-order Taylor expansion. The resulting variance estimator, denoted by  $\tilde{v}_{JF}$ , is the linearization jackknife variance estimator studied by Yung and Rao (2000). Similarly, the shortcut jackknife variance estimator  $v_{JS}$  is asymptotically equivalent to a variance estimator of  $V_p(\hat{Y}_{\text{CAL}} | \mathbf{r})$  obtained by treating the nonresponse weighting adjustment factors  $g_{hi}$  as constants. We denote this variance estimator by  $\tilde{v}_{JS}$ . The quantity  $D$  can thus be approximated by  $\tilde{D} = \tilde{v}_{JS} - \tilde{v}_{JF}$ . For this approximation to be valid, we assume the number of respondents to be large.

Noting that  $\text{Bias}(v_{JF}) = E_{pq}(v_{JF}) - V(\hat{Y}_{\text{CAL}}) \approx 0$ , it follows that the bias of  $v_{JS}$ ,  $\text{Bias}(v_{JS}) = E_{pq}(v_{JS}) - V(\hat{Y}_{\text{CAL}})$ , can be approximated by  $E_{pq}(D) \approx E_{pq}(\tilde{D})$ . Let  $v(y)$  denote the variance estimator of the complete data estimator (1.1). Using a first-order Taylor expansion, it can be shown that an estimator of  $V_p(\hat{Y}_{\text{CAL}} | \mathbf{r})$  is given by

$$\tilde{v}_{JF} = v(\hat{\xi}) \quad (2.4)$$

where

$$\hat{\xi}_{hi} = \mathbf{x}_{hi}' \hat{\mathbf{B}}_r + g_{hi} r_{hi} e_{hi},$$

with  $e_{hi} = (y_{hi} - \mathbf{x}_{hi}' \hat{\mathbf{B}}_r)$  and  $\hat{\mathbf{B}}_r = \hat{\mathbf{T}}_r^{-1} \sum_{(hi) \in s} d_{hi} r_{hi} \mathbf{z}_{hi} y_{hi}$ . On the other hand, treating the  $g_{hi}$ 's as constants implies that  $\hat{Y}_{\text{CAL}}$  is linear in the design weights  $d_{hi}$ . It follows that  $\tilde{v}_{JS}$  is given by

$$\tilde{v}_{JS} = v(\Psi), \quad (2.5)$$

where  $\Psi_{hi} = g_{hi} r_{hi} y_{hi}$ .

For example, for either a fixed size or a random size sampling design, a possible variance estimator is

$$\tilde{v}_{JF} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \hat{\xi}_i \hat{\xi}_j,$$

where  $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j) / \pi_j \pi_i \pi_j$  and  $\pi_{ij}$  is the second-order inclusion probability of units  $i$  and  $j$ . Note that  $\pi_{ii} = \pi_i$ . Similarly, we have

$$\tilde{v}_{JS} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \psi_i \psi_j.$$

### 3. Bias of $v_{JS}$ in some special cases

#### 3.1 Simple random sampling without replacement

In this section, we assume that the sample  $s$  has been selected according to simple random sampling without replacement. We also assume that the sampling fraction  $n/N$  is negligible and that the number of respondents  $r$  is large. Finally, we assume a single weighting class. Although the above situation is not realistic in practice, it provides some insight into the asymptotic bias of  $v_{JS}$ .

In the case of the ratio adjusted estimator (1.5), we can show that  $\tilde{D}$  is approximately given by

$$\begin{aligned} \tilde{D} = & \frac{N^2}{r} \left(1 - \frac{r}{n}\right) \left\{ \left( \frac{\bar{x}}{\bar{x}_r} \right)^2 (s_{1r}^2 - s_{cr}^2) \right. \\ & + 2 \left( \frac{\bar{x}}{\bar{x}_r} \right) \hat{R}_r \left[ \left( \frac{\bar{x}}{\bar{x}_r} \right) - 1 \right] \frac{s_{cxy}}{n} \\ & \left. + \hat{R}_r^2 \left[ \left( \frac{\bar{x}}{\bar{x}_r} \right)^2 s_{xr}^2 - s_x^2 \right] + \left( \frac{\bar{x}}{\bar{x}_r} \right)^2 \bar{y}_r^2 \right\}, \quad (3.1) \end{aligned}$$

where  $(\bar{x}_r, \bar{y}_r) = 1/r \sum_{i \in s} r_i (x_i, y_i)$  denote the mean of the respondents for variable  $x$  and  $y$  respectively and  $r$  is the number of respondents,  $\hat{R}_r = \bar{y}_r / \bar{x}_r$ ,  $s_{xr}^2 = 1/(r-1) \sum_{i \in s} r_i (x_i - \bar{x}_r)^2$ ,  $s_x^2 = 1/(n-1) \sum_{i \in s} (x_i - \bar{x})^2$  with  $\bar{x} = 1/n \sum_{i \in s} x_i$ ,  $s_{cr}^2 = 1/(r-1) \sum_{i \in s} r_i (y_i - \hat{R}_r x_i)^2$  and  $s_{cxy} = 1/(r-1) \sum_{i \in s} r_i (y_i - \hat{R}_r x_i) x_i$ . If we further assume that all units have equal response probabilities (*i.e.*, a uniform response mechanism), we have  $\bar{x} / \bar{x}_r \xrightarrow{p} 1$  and  $s_{xr}^2 / s_x^2 \xrightarrow{p} 1$ . In this case, the asymptotic bias of  $v_{JS}$  is given by

$$\begin{aligned} \text{Bias}(v_{JS}) & \approx E_{pq}(\tilde{D}) \\ & \approx \frac{N^2}{E_{pq}(r)} \left(1 - E_{pq}\left(\frac{r}{n}\right)\right) \\ & S_y^2 \left( \frac{1}{CV(y)^2} + 2 \frac{CV(x)}{CV(y)} \rho_{xy} - \frac{CV(x)^2}{CV(y)^2} \right), \quad (3.2) \end{aligned}$$

where  $CV(x) = S_x / \bar{X}$  and  $CV(y) = S_y / \bar{Y}$  denote the population coefficients of variation for variables  $x$  and  $y$ , respectively with  $S_y^2 = 1/(N-1) \sum_{i \in U} (y_i - \bar{Y})^2$  and  $\bar{Y} = 1/N \sum_{i \in U} y_i$ ,  $S_x^2$  and  $\bar{X}$  are defined similarly, and  $\rho_{xy}$  denotes the finite population coefficient of correlation for variables  $x$  and  $y$ . From (3.2), it follows that the asymptotic bias of  $v_{JS}$  is nonnegative if and only if

$$B_0 < \frac{\bar{Y}}{2} \left( \frac{1 + CV(x)^2}{CV(x)^2} \right), \quad (3.3)$$

provided  $0 < E_{pq}(r/n) < 1$ , where  $B_0 = \bar{Y} - B_1 \bar{X}$  is the finite population intercept of the least squares line when regressing  $y$  on  $x$  with

$$B_1 = \frac{\sum_{i \in U} (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i \in U} (x_i - \bar{X})^2}.$$

From (3.2), it is clear that the bias of  $v_{JS}$  increases if (i) the expected response rate  $E_{pq}(r/n)$  decreases; (ii)  $\rho_{xy}$  increases; (iii)  $CV(y)$  decreases; or (iv)  $CV(x)$  increases. Also, it follows from (3.3) that  $v_{JS}$  overestimates the true variance when the intercept  $B_0$  is not too large. Table 1 illustrates the relationship between  $CV(x)$  and the condition in (3.3). For example, when  $CV(x) = 0$ ,  $v_{JS}$  always overestimates the true variance since, in this case, the condition (3.3) reduces to  $B_0 < \infty$ , which is always satisfied. This result is not surprising because when  $CV(x) = 0$ , the  $x$ -values are all equal and the ratio adjusted estimator (1.5) is identical to the count adjusted estimator (1.4). As we discuss below,  $v_{JS}$  always overestimates the true variance in this case. When  $CV(x)$  is large (*e.g.*,  $CV(x) = 2$ ),  $v_{JS}$  overestimates the true variance if and only if  $B_0 < 0.625 \bar{Y}$ . The latter condition is satisfied if the intercept is not “too far” from the origin. Therefore, if the relationship between  $y$  and  $x$  goes through the origin (*i.e.*, if the ratio model holds), the shortcut variance estimator will overestimate the true variance. However, if the ratio adjusted estimator is used when the ratio model does not hold, such as when  $B_0 \geq 0.625 \bar{Y}$ , the shortcut variance estimator  $v_{JS}$  will underestimate the true variance. In conclusion, we can expect  $v_{JS}$  to overestimate the true variance when a ratio adjustment procedure is used unless the ratio model is highly misspecified for the data at hand, which could happen, for example, if the variables  $y$  and  $x$  are negatively correlated.

**Table 1**  
Relationship between  $CV(x)$  and the condition in (3.3)

$CV(x)$	$\frac{\bar{Y}}{2} \left( \frac{1 + CV(x)^2}{CV(x)^2} \right)$
0	$\infty$
0.1	$50.5 \bar{Y}$
0.5	$2.5 \bar{Y}$
1	$2 \bar{Y}$
1.5	$0.722 \bar{Y}$
2	$0.625 \bar{Y}$

Turning to the count adjusted estimator (1.4), we let  $x_i = 1$  for all  $i$  in (3.1) and obtain

$$\tilde{D} = \frac{N^2}{r} \left(1 - \frac{r}{n}\right) \bar{y}_r^2. \quad (3.4)$$



It follows from (3.4) that the relative bias of  $v_{JS}$ ,  $RB(v_{JS}) = \text{Bias}(v_{JS})/V(\hat{Y}_{CAL})$ , can be approximated by  $E_{pq}(R\tilde{D})$  where  $R\tilde{D} = \tilde{D}/\tilde{v}_{JF}$ . Under a uniform nonresponse mechanism, straightforward algebra leads to

$$RB(v_{JS}) \approx E_{pq}(R\tilde{D}) \approx \left(1 - E_{pq}\left(\frac{r}{n}\right)\right) \frac{1}{CV(y)^2}. \quad (3.5)$$

The expression (3.5) shows that, in the case of the count adjusted estimator (1.4),  $v_{JS}$  always overestimates the true variance. The magnitude of the overestimation increases as the expected response rate  $E_{pq}(r/n)$  decreases or when  $CV(y)$  decreases. For example, if the expected response rate is equal to 70% and  $CV(y) = 1$ , we have  $E_{pq}(R\tilde{D}) = 1.3$  so the shortcut jackknife variance estimator,  $v_{JS}$ , is on average 30% larger than the true variance of  $\hat{Y}_{CAL}$ . On the other hand, if the response rate is equal to 70% and  $CV(y) = 0.5$ , we have  $E_{pq}(R\tilde{D}) = 5.3$ , in which case the overestimation is considerable.

Finally, we turn to the case of the simple linear regression adjusted estimator (1.6). Under a uniform nonresponse mechanism, it can be shown that the asymptotic bias of  $v_{JS}$  is given by

$$\begin{aligned} \text{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \frac{N^2}{E_{pq}(r)} \left(1 - E_{pq}\left(\frac{r}{n}\right)\right) \\ &\quad S_y^2 \left(\frac{1}{CV(y)^2} + \rho_{xy}^2\right) \geq 0. \end{aligned} \quad (3.6)$$

From (3.6), it follows that  $v_{JS}$  always overestimates the true variance in the case of the simple linear regression adjusted estimator (1.6). The bias (3.6) increases if (i) the expected response rate decreases; (ii)  $\rho_{xy}^2$  increases; or (iii)  $CV(y)$  decreases.

### 3.2 Stratified simple random sampling: Weighting classes are identical to strata

In this section, we assume that the weighting classes coincide with the original design strata. This situation is not uncommon in practice, especially in business surveys. If the strata are such that the units within stratum have approximately equal response propensities (*i.e.*, uniform response within stratum), expressions for the bias of  $v_{JS}$  are readily obtained from expressions (3.2), (3.4) and (3.6).

For the ratio adjusted estimator, expression (3.2) can be readily extended to the case of stratified simple random sampling to obtain

$$\begin{aligned} \text{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^L \frac{N_h^2}{E_{pq}(r_h)} \left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right) \\ &\quad S_{yh}^2 \left(\frac{1}{CV_h(y)^2} + 2 \frac{CV_h(x)}{CV_h(y)} \rho_{hxy} - \frac{CV_h(x)^2}{CV_h(y)^2}\right), \end{aligned} \quad (3.7)$$

where the quantities  $r_h$ ,  $CV_h(x)$ ,  $CV_h(y)$ ,  $S_{yh}^2$  and  $\rho_{hxy}$  correspond to  $r$ ,  $CV(x)$ ,  $CV(y)$ ,  $S_y^2$  and  $\rho_{xy}$  computed in each stratum.

For the count adjusted estimator, expression (3.4) can be readily extended to the case of stratified simple random sampling to obtain

$$\begin{aligned} \text{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^L \frac{N_h^2}{E_{pq}(r_h)} \left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right) \frac{S_{yh}^2}{CV_h(y)^2}. \end{aligned} \quad (3.8)$$

Finally, for the simple linear regression adjusted estimator, expression (3.6) can be readily extended to the case of stratified simple random sampling to obtain

$$\begin{aligned} \text{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^L \frac{N_h^2}{E_{pq}(r_h)} \left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right) \\ &\quad S_{yh}^2 \left(\frac{1}{CV_h(y)^2} + \rho_{hxy}^2\right). \end{aligned} \quad (3.9)$$

From the expressions (3.7)-(3.9), it follows that the use of the shortcut jackknife variance estimator requires some caution. Indeed, even if the bias of the shortcut jackknife variance estimator is small in each stratum, they might sum up to a considerable bias at the population level if the biases are in the same direction.

## 4. Simulation study

A simulation study was performed to compare the statistical properties of the shortcut and the full jackknife variance estimators under varying conditions. Five different stratified populations of 30,000 units each with two variables were generated. First, the  $x$ -values were generated from a Gamma distribution with parameters  $\alpha$  and  $\lambda$ . Then given the  $x$ -values, the  $y$ -values were generated according to the following model:

$$y_{hi} = \beta_0 + \beta_1 x_{hi} + \varepsilon_{hi},$$

where  $\varepsilon_{hi} \sim N(0, \sigma_{\varepsilon h}^2)$ . The variance and  $\sigma_{\varepsilon h}^2$  was set such that the coefficient of correlation (denoted  $\rho_{xy}$ ) between  $x_{hi}$  and  $y_{hi}$  is equal to 0.7 in all the populations. Each population was stratified into three strata, each with 10,000 units. The parameters of the simulated populations appear in Table 2.

Population 1 fits the ratio model very well with an intercept of zero in all strata. Population 2 has a non-negligible intercept term in all three strata. Population 3 is a mix of populations 1 and 2, where the ratio model fits well for strata 2 and 3 but not for stratum 1. Population 4 is

similar to population 1 except units in strata 1 and 2 have a 70% chance of reporting a zero. This population is intended to mimic the situation of the Annual Capital Expenditures Survey (ACES) of the U.S. Census Bureau, which provided the motivation for this research. The ACES employs a shortcut jackknife variance estimator that, empirically, has been shown to be close to the full jackknife variance estimates. Its population is characterized with many zeros for capital expenditures in the majority of sampled small and medium businesses, with the majority of the reported expenditures being provided by large businesses. Population 5 was generated to show that the shortcut estimator for the ratio adjusted estimator can actually have a negative bias when the ratio model is misspecified (demonstrated in expression (3.3) for a simple random sample). For this population, the intercept term is highly significant in all strata.

**Table 2**  
**Population parameters**

Population	$\beta_0$			$\beta_1$			$\alpha$	$\lambda$	CV(x)	CV(y)
	(Within Stratum)			(Within Stratum)						
	1	2	3	1	2	3				
1	0	0	0	2	4	6	4	5	50%	76%
2	120	240	360	2	4	6	4	5	50%	44%
3	120	0	0	2	4	6	4	5	50%	51%
4	0	0	0	2	4	6	4	5	50%	134%
5	50	200	300	0.5	1	2	4	5	200%	63%

From each population, 5,000 stratified simple random samples of size 300 (100 units per stratum) were drawn. In each sample, nonresponse was generated using a uniform response mechanism within each stratum with probabilities of response equal to 60% in stratum 1, 70% in stratum 2 and 90% in stratum 3. This response pattern is not uncommon in business surveys where more follow-up is performed for the medium and large size units (strata 2 and 3).

In each sample, both the count adjusted and the ratio adjusted estimators, given respectively by (1.4) and (1.5), were calculated using the strata as weighting classes. The variance of the point estimators was estimated by  $v_{JF}$  and  $v_{JS}$ , given respectively by (2.2) and (2.3). As a measure of the bias of a variance estimator  $v$ , we used the Monte Carlo percent relative bias given by

$$RB_{MC}(v) = \frac{1}{5,000} \sum_{t=1}^{5,000} \frac{v^{(t)} - MSE_{MC}(\hat{Y}_{CAL})}{MSE_{MC}(\hat{Y}_{CAL})} \times 100,$$

where  $v^{(t)}$  is the variance estimate obtained from the  $t^{th}$  sample, and  $MSE_{MC}(\hat{Y}_{CAL})$  is the Monte Carlo Mean Squared Error (MSE) defined by

$$MSE_{MC}(\hat{Y}_{CAL}) = \frac{1}{50,000} \sum_{t=1}^{50,000} (\hat{Y}_{CAL}^{(t)} - Y)^2,$$

where  $\hat{Y}_{CAL}^{(t)}$  is the (ratio or count adjusted) estimate of  $Y$  for the  $t^{th}$  sample. Table 3 shows the Monte Carlo percent relative bias for both the count adjusted and the ratio adjusted estimators.

**Table 3**  
**Monte Carlo percent relative bias for the shortcut and full jackknife variance estimators**

Population	Count adjusted estimator		Ratio adjusted estimator	
	$RB_{MC}(v_{JS})$	$RB_{MC}(v_{JF})$	$RB_{MC}(v_{JS})$	$RB_{MC}(v_{JF})$
1	57.3%	1.1%	80.5%	-0.3%
2	877.1%	0.4%	364.7%	0.5%
3	220.7%	0.6%	185.9%	-0.2%
4	21.6%	0.6%	29.1%	1.4%
5	266.4%	0.2%	-67.2%	5.0%

As expected, the shortcut estimator overestimates the Monte Carlo MSE for the count adjusted estimator for all populations. The overestimation varies from approximately 20% in population 4 to over 800% in population 2. From expression (3.8), we see that the bias of  $v_{JS}$  depends on the response rate and  $\bar{y}_h^2$ . Population 2 has a large intercept term which increases  $CV_h(y)$  in all strata, which in turn increases the bias of  $v_{JS}$ . Population 3 is similar to population 2 except only the first stratum has a large intercept term. As expected, the bias of  $v_{JS}$  in this population is between those of populations 1 and 2. Population 4 is the one generated to mimic the ACES population with some units' values replaced by zero in strata 1 and 2. The Monte Carlo relative bias of 21.6% is, for the most part, coming from the third stratum where no units have been replaced with zero (this can be seen using expression (3.8)). In comparison, for all five populations the full jackknife variance estimator is tracking the Monte Carlo MSE very well with absolute relative biases less than 1.1%.

Turning to the ratio adjusted estimator, we see that the full jackknife variance estimator again tracks the Monte Carlo MSE relatively well for all populations with absolute relative biases less than 5%. The shortcut estimator, on the other hand, has relative biases varying from -67% to 364%. Looking at expression (3.7), we see that for a fixed response rate the bias depends on the  $CV_h(y)$ ,  $CV_h(x)$  and  $\rho_{hxy}$ . Due to the large intercept terms in the second population,  $\bar{y}_h$  are large and the corresponding  $CV_h(y)$  are smaller than in the other populations. Thus, the last term in expression (3.7) is quite large and the resulting relative bias of  $v_{JS}$  is also large. This is also seen for population 3 except to a lesser extent since only the first stratum has an intercept term. The opposite effect is seen in population 4, where the introduction of zeros has significantly increased  $CV_h(y)$  which has in turn reduced the Monte Carlo percent relative bias of the shortcut estimator.



Additional simulations were performed using the some of the populations described in Table 2 but with varying response rates. The results are not presented here as they were as expected. That is, the bias of the shortcut estimator decreased as the response rate increased (with all the other parameters remaining fixed). The full jackknife estimator continued to track the Monte Carlo MSE very well.

## 5. Conclusion

In this paper, we evaluated both theoretically and empirically a shortcut jackknife variance estimator that does not re-calculate the nonresponse adjustment factors within each jackknife replicate, specifically considering three different nonresponse weighting adjustment procedures. We showed in the context of stratified simple random sampling that the shortcut jackknife variance estimator tends to overestimate the true variance of the estimators. In the context of the ratio adjustment procedure, however, the shortcut jackknife variance estimator may underestimate the true variance if the ratio model is not appropriate for the data at hand.

One justification for the use of a shortcut procedure in a replicate variance estimation method is to save time and computing resources. If these are truly issues and the program has consistently high unit response rates in all weighting cells, then while there are clearly theoretical advantages to replicating the weight adjustment procedure, there may be little or no practical advantage. Having said that, the conditions for “practical” equivalence between the full and shortcut procedure variance estimators are extremely restrictive, and we have demonstrated that small changes in underlying data conditions can easily violate these conditions. If computational concerns with a full jackknife are truly an issue, then the authors recommend the linearization jackknife variance estimation approach which has the same asymptotic properties as the full jackknife, but is computationally quick and computer overhead “free” (in terms of replicate storage). See Thompson and Yung (2006) for expressions for the linearization jackknife variance estimator for both the count and ratio adjusted estimators. Given these viable alternatives, we recommend against the use of a shortcut procedure variance estimator.

## Acknowledgements

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. The

authors would like to thank the Associate Editor, two anonymous referees, Samson Adeshiyan, Patrick Cantwell, Carol Caldwell, Michael Hidirolou, Rita Petroni, Mark Sands, and Jun Shao for their useful comments on earlier versions of this paper. Work of David Haziza was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

## References

- Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society B*, 67, 445-458.
- Brick, M.J., and Montaquila, J.M. (2009). Nonresponse and weighting. In the *Handbook of Statistics, Sample Surveys: Design, Methods and Applications*, (Eds., C.R. Rao and D. Pfeffermann), 29A, 163-185.
- Da Silva, D.N., and Opsomer, J.D. (2006). A kernel smoothing method to adjust for unit nonresponse in sample surveys. *Canadian Journal of Statistics*, 34, 563-579.
- Deville, J.-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des Journées de Méthodologie Statistique*, Insee.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Eltinge, J.L., and Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.
- Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429-440.
- Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.
- Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501-514.
- Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.
- Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- Ozcokun, L., Thompson, K.J. and Williams, Q. (2005). Investigation of balanced repeated replication (BRR) variance estimation for the Survey of Residential Alterations and Repairs (SORAR). *Proceedings of the Federal Committee on Statistical Methodology*, Office of Management and Budget.
- Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.
- Sautory, O. (2003). CALMAR 2: A new version of the CALMAR calibration adjustment program. *Proceedings: Symposium 2003, Challenges in Survey Taking for the Next Decade*, Ottawa, Canada.

- Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 93, 254-265.
- Thompson, K.J. (2005). An empirical investigation into the effects of replicate reweighting on variance estimates for the annual capital expenditures survey. *Proceedings of the Federal Committee on Statistical Methodology*, Office of Management and Budget.
- Thompson, K.J., and Yung, W. (2006). To Replicate (A weight adjustment procedure) or not to replicate? An analysis of the variance estimation effects of a shortcut procedure using the stratified jackknife. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3772-3779.
- Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics*, 20, 1-18.
- Yung, W., and Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95, 903-915.





# A comparison of variance estimators for poststratification to estimated control totals

Jill A. Dever and Richard Valliant<sup>1</sup>

## Abstract

Calibration techniques, such as poststratification, use auxiliary information to improve the efficiency of survey estimates. The control totals, to which sample weights are poststratified (or calibrated), are assumed to be population values. Often, however, the controls are estimated from other surveys. Many researchers apply traditional poststratification variance estimators to situations where the control totals are estimated, thus assuming that any additional sampling variance associated with these controls is negligible. The goal of the research presented here is to evaluate variance estimators for stratified, multi-stage designs under estimated-control (EC) poststratification using design-unbiased controls. We compare the theoretical and empirical properties of linearization and jackknife variance estimators for a poststratified estimator of a population total. Illustrations are given of the effects on variances from different levels of precision in the estimated controls. Our research suggests (i) traditional variance estimators can seriously underestimate the theoretical variance, and (ii) two EC poststratification variance estimators can mitigate the negative bias.

Key Words: Estimated-control poststratification; Sampling frame coverage bias; Survey-estimated control totals.

## 1. Introduction

Poststratified estimators, and other calibration estimators, are used in many types of surveys to reduce variances or to correct for frame deficiencies. Specific examples include large U.S. government surveys, such as the Consumer Expenditure Survey (see, *e.g.*, Jayasuriya and Valliant 1996); surveys of specialized populations, such as the U.S. Department of Defense Survey of Health Related Behaviors among Military Personnel (Bray, Hourani, Rae, Dever, Brown, Vincus, Pemberton, Marsden, Faulkner and Vandermaas-Peeler 2003); and a myriad of surveys outside the U.S. including the Canadian Retail Trade Survey (see, *e.g.*, Hidioglou and Patak 2006), the Swedish Labour Force Survey (Mirza and Hörmgren 2002), and the British Household Panel Survey (Taylor, Brice, Buck and Prentice-Lane 2007).

Calibration estimators, such as those generated under poststratification, are used to minimize errors associated with incomplete sampling frames (*i.e.*, undercoverage) and with sampling and nonresponse (see, *e.g.*, Särndal, Swensson and Wretman 1992; Lessler and Kalsbeek 1992; Kott 2006). For example, estimates from the Behavioral Risk Factor Surveillance System (BRFSS), a nationwide random-digit-dial (RDD) telephone survey conducted by the U.S. Centers for Disease Control and Prevention (CDC), are poststratified to counts that include households with and without landline telephone service (Centers for Disease Control and Prevention 2006). The decrease in the errors is linked to the association of the population control totals with the frame

undercoverage, patterns of non-ignorable nonresponse, and the variable of interest (Kim, Li and Valliant 2007).

When relevant population controls do not exist, many researchers use survey-estimated control totals, and apply traditional variance formulae as if the controls were known without error. For example, Nadimpalli, Judkins and Chu (2004) adjusted weights for the *2003 National Survey of Parents and Youth* to the number of U.S. households with children ages 9-18 estimated from the *Current Population Survey* (CPS) using a ratio-raking algorithm ([www.census.gov/cps](http://www.census.gov/cps)). Estimates of how people in the U.S. spend their time can be calculated from *The American Time Use Survey* using weights that have been poststratified to projected estimates from the U.S. decennial Census (Killion 2006). More recently, researchers at the Pew Research Centers calibrated weights for a set of 2008 U.S. presidential pre-election surveys to population estimates from the March 2007 CPS, as well as to estimates on telephone usage patterns from the July-December 2007 *National Health Interview Survey* (Keeter, Dimock and Christian 2008).

The goal of our research is to develop and evaluate variance estimators for point estimates with weights that contain a poststratification adjustment to a set of survey-estimated control totals. We label the methodology which properly accounts for the estimated controls as *estimated-control (EC) poststratification*. In this paper, we focus specifically on the EC poststratified (ECPs) estimator of a population total for data collected from a stratified, multi-stage design, where the first-stage sampling units are selected with *replacement*. The remainder of this section gives a brief review of weight calibration and poststratification. Section 2

1. Jill A. Dever, RTI International. Email: [jdever@rti.org](mailto:jdever@rti.org); Richard Valliant, Survey Research Center, University of Michigan and Joint Program in Survey Methodology, University of Maryland. Email: [rvalliant@survey.umd.edu](mailto:rvalliant@survey.umd.edu).



contains an explicit definition of the ECPS estimator under study, followed in Section 3 by an evaluation of the bias properties. Through a theoretical evaluation (Section 4) and a simulation study, we compare variance estimators developed for the ECPS estimator with a variance estimator chosen under the naïve “population control total” assumption. Both linearization and replication variance estimators are examined in our research. We provide illustrations on the effects of different levels of precision in the estimated controls on the variance estimates. The specifications for the simulation study are detailed in Section 5, followed by a summary of the results (Section 6). We conclude the paper with a brief summary and an overview of future research in this area.

*Calibration estimators* (Deville and Särndal 1992), such as a poststratified estimator of a population total, borrow strength from auxiliary information to improve the efficiency of survey estimates over simpler weighting methods. When the auxiliary variables are (linearly) related to the set of key survey variables, calibration estimators can be very efficient.

The general form of a *traditional* or *fixed-control* calibration estimator is best described as an expansion estimator or “linear weighting” estimator as discussed in Estevao and Särndal (2000). Define  $s$  to be the set of sample elements from a probability sample, and  $d_k = 1/\pi_k$  to be the design weight for element  $k$  such that  $\pi_k = \Pr(k \in s)$ . An estimated population total of a variable  $y$  is  $\hat{t}_y = \sum_{k \in s} w_k y_k$ , where the calibration weight ( $w_k = a_k d_k$ ) for the  $k^{\text{th}}$  element defined as a function of the design weight,  $d_k$ , and a calibration-adjustment factor,  $a_k$ , also known as a  $g$ -weight (Särndal *et al.* 1992). The calibration weights are calculated by minimizing a specified function that measures the distance between the design and calibration weights subject to a set of constraints defined as:

$$\mathbf{t}_{U_x} = \hat{\mathbf{t}}_{A_x} \quad (1)$$

where  $\mathbf{t}_{U_x} = \sum_{k \in U} \mathbf{x}_k$ , the vector of population controls (counts) corresponding to the  $G$  ( $G \geq 1$ ) auxiliary variables;  $\hat{\mathbf{t}}_x = \sum_{k \in s} w_k \mathbf{x}_k$ , the estimated population controls corresponding to the components of  $\mathbf{t}_{U_x}$ ; and  $\mathbf{x}_k$  is a vector of length  $G$  containing auxiliary or benchmark variable values for element  $k$ . Note that  $\mathbf{x}_k$  may contain ones and zeros to indicate the presence or absence of a certain characteristic (e.g., age 18-25), or larger values (e.g., number of children). An example of such a calibration system is the generalized least squares (or chi-square) distance function  $\sum_{k \in s} (w_k - d_k)^2 / c_k d_k$  that is minimized subject to the constraints in (1). This system generates a closed-form solution called the generalized regression estimator (GREG) for  $c_k = 1$  (Deville and Särndal 1992). The poststratified estimator is a special case of the GREG.

Variance estimation techniques for the poststratified estimator, and more generally for the GREG, have been widely studied. Binder (1995) demonstrates techniques used to calculate a *Taylor linearization* variance estimator for the GREG. Additional references for the linearization variance estimator under poststratification (and calibration more generally) include Deville, Särndal and Sautory (1993), Demnati and Rao (2004), and Hidiogrou and Patak (2006). Särndal, Swensson and Wretman (1989) developed an approximate linearization variance for the GREG of a population total as a function of the population residuals from a specified model and the design weights ( $d_k$ ). Valliant (1993) and Yung and Rao (1996) modified the residual-based variance estimator by multiplying the sample residuals by the calibration weights  $w_k (= a_k d_k)$ . They demonstrated that this revised estimator, created by linearizing the associated jackknife, reduced the bias associated with the original formula. This variance estimator is also discussed in Särndal *et al.* (1992), Stukel, Hidiogrou and Särndal (1996), and in Chapter 11 of Särndal and Lundström (2005). Properties of replication variance estimators (i.e., jackknife and BRR) have been examined in, for example, Valliant (1993), Rust and Rao (1996), Canty and Davison (1999), Théberge (1999), Rao and Shao (1999), Yung and Rao (1996; 2000), and Kott (2006).

An assumption in the articles above is that the control totals, to which the auxiliary sample estimates are adjusted, are either true population values known without error, or are taken from an independent, highly precise survey that is much larger than the survey requiring calibration. In some cases, however, these controls are estimated from other surveys with non-negligible sampling variances. For example, there are efforts to calibrate Web panel surveys to separate, higher-quality reference surveys that are not much larger than the panel surveys themselves (e.g., Krotki 2007; Terhanian, Bremer, Smith and Thomas 2000).

Many researchers apply formulae developed for traditional poststratification even though the controls have been estimated. The tacit assumption is that any additional error (variance and bias) associated with these controls is negligible and can be ignored. Currently, the validity of this assumption can not be checked until a complete picture of EC poststratification has been developed.

## 2. The estimated-control poststratified estimator

To facilitate our discussion of the estimated-control poststratified estimator, we label the survey requiring poststratification as the *analytic survey* and the source of the control totals as the *benchmark survey*. In practice, more than one benchmark survey may be tapped for the control totals. However, we will assume only one benchmark

survey for the theoretical development so that control total variances and covariances are estimable.

Let  $U$  represent the finite target population containing  $N$  elements and  $t_y = \sum_{k \in U} y_k$  represent the population total of interest for a variable  $y$ . Let  $s_A$  represent a random sample of size  $n_A$  from the frame  $U_A$  for the analytic survey. A random sample  $s_B$  of size  $n_B$  is selected for the benchmark survey from the corresponding sampling frame  $U_B$ . We allow the possibility that each of the frames,  $U_A$  and  $U_B$ , do not completely cover the target population  $U$ . However, coverage is treated as a random event so that all elements in the target population have a positive probability of being covered by either the analytic or the benchmark survey frame.

As a convention throughout the paper, an “A” subscript signifies an association with the analytic survey such as a sample design parameter or an estimate. A “B” subscript identifies the benchmark survey quantities. These subscripts are absent from the parameters associated with the population of interest, *i.e.*,  $t_y$ .

For the stratified, multi-stage design assumed for the analytic survey,  $m_{Ah}$  ( $m_{Ah} \geq 2$ ) primary sampling units (PSUs), indexed by  $i$ , are selected *with replacement* from a total of  $M_{Ah}$  PSUs in the  $h^{\text{th}}$  design stratum ( $h = 1, \dots, H$  with  $H \geq 2$ ). We assume that  $n_{Ahi}$  elements, each indexed by  $k$ , are selected from  $N_{Ahi}$  in PSU  $hi$  in such a way that an unbiased estimate of the PSU total can be made. The design weight,  $d_k$ , is calculated as the inverse of the unconditional inclusion probability for  $k \in s_{Ahi}$ , the set of analytic survey elements within the  $hi^{\text{th}}$  PSU. Thus,  $n_A$ , the size of the analytic survey sample, is calculated as  $n_A = \sum_{h=1}^H \sum_{i=1}^{m_{Ah}} n_{Ahi}$ . Elements for the benchmark survey are randomly drawn from the corresponding sampling frame; no explicit specifications are made for the random sampling method.

Poststratification can be used to correct for sampling and coverage errors. Therefore, we allow undercoverage in the analytic-survey, as well as, the benchmark-survey sampling frames. Additionally, we do not consider the effects of nonresponse.

Suppose that the population  $U$  can be divided into  $g = 1, \dots, G$  mutually exclusive and exhaustive poststrata. When the population count of elements,  $N_g$ , is known for each poststratum, the traditional poststratified estimator of a total for  $y$  is defined as

$$\hat{t}_{yPS} = \sum_{g=1}^G N_g \frac{\hat{t}_{Ay g}}{\hat{N}_{Ag}}, \quad (2)$$

where  $y_k$  is the value of the analysis variable  $y$  for element  $k$ ;  $\hat{t}_{Ay g} = \sum_{k \in s_A} \delta_{gk} d_k y_k$ , the total of  $y$  in poststratum  $g$  estimated from the analytic survey data;  $\hat{N}_{Ag} = \sum_{k \in s_A} \delta_{gk} d_k$ , the analytic survey estimated total in poststratum  $g$ ; and

$\delta_{gk} = 1$  indicates membership in the  $g^{\text{th}}$  poststratum and zero otherwise. Note that  $\hat{t}_{Ay g}$  may also be expressed as  $\hat{t}_{Ay g} = \sum_{k \in s_{Ag}} d_k y_k$ , where  $s_{Ag}$  indicates the set of analytic survey elements in poststratum  $g$ . The “hat” notation in the expression above is used to distinguish a population estimator (*e.g.*,  $\hat{N}_{Ag}$ ) from the known population parameter (*e.g.*,  $N_g$ ). If the count of elements in poststratum  $g$  is estimated by setting  $y_k = 1$  in the formula for  $\hat{t}_{Ay g}$ , then  $\hat{t}_{yPS}$  equals  $N_g$ . In this sense,  $\hat{t}_{yPS}$  is poststratified to the population counts  $N_1, \dots, N_G$ .

In certain situations, however, the population counts are not available and must be estimated from a benchmark survey. Define the ECPS estimator of a population total of a variable  $y$  as

$$\hat{t}_{yP} = \sum_{g=1}^G \hat{N}_{Bg} \frac{\hat{t}_{Ay g}}{\hat{N}_{Ag}}. \quad (3)$$

The number of population elements in the  $g^{\text{th}}$  poststratum ( $g = 1, \dots, G$ ) estimated from the benchmark survey is denoted as  $\hat{N}_{Bg} = \sum_{l \in s_{Bg}} w_l$ , where  $s_{Bg}$  is the set of sample elements in poststratum  $g$  from the benchmark survey and  $w_l$  is the weight associated with the  $l^{\text{th}}$  element. The calibration-adjustment factors applied to the analytic survey design weights for  $\hat{t}_{yP}$  are calculated as  $a_k = \hat{N}_{Bg} / \hat{N}_{Ag}$  for  $k \in s_{Ag}$ .

Relating the poststratified estimators to the calibration system discussed in the previous section,  $\hat{\mathbf{t}}_{Ax}$  is a  $G$ -length vector of estimated population counts for each poststratum such that  $\hat{\mathbf{t}}_{Ax} = (\hat{t}_{Ax1}, \dots, \hat{t}_{AxG})'$ , where  $\hat{t}_{Axg} \equiv \hat{N}_{Ag} = \sum_{k \in s_A} d_k \delta_{gk}$  and  $x_k \equiv \delta_{gk} = 1$  if the element  $k$  is a member of the  $g^{\text{th}}$  poststratum and 0 otherwise. The vector  $\mathbf{t}_{Ux}$  corresponds either to  $\mathbf{N} = (N_1, \dots, N_G)'$  for the  $\hat{t}_{yPS}$  estimator given in (2), or to  $\hat{\mathbf{N}}_B = (\hat{N}_{B1}, \dots, \hat{N}_{BG})'$ , a  $G \times 1$  vector of benchmark control estimates, for the  $\hat{t}_{yP}$  estimator given in (3).

The estimator  $\hat{t}_{yP}$  can be expressed in matrix notation as  $\hat{t}_{yP} = \hat{\mathbf{N}}_B' \hat{\mathbf{Y}}_A$  where  $\hat{\mathbf{Y}}_A = (\hat{\mathbf{N}}_A)^{-1} \hat{\mathbf{t}}_{Ay}$ , a  $G \times 1$  vector of analytic survey estimates of the form  $\hat{\mathbf{Y}}_1 = [\hat{t}_{11} / \hat{N}_{A1}, \dots, \hat{t}_{AG} / \hat{N}_{AG}]'$ ;  $\hat{\mathbf{N}}_A = \text{diag}(\hat{N}_{A1}, \dots, \hat{N}_{AG})$ , a diagonal matrix of poststratum totals estimated from the analytic survey; and  $\hat{\mathbf{t}}_{Ay} = [\hat{t}_{A1}, \dots, \hat{t}_{AG}]'$  is a  $G \times 1$  vector of poststratum totals for the outcome variable estimated from the analytic survey. The remaining variables associated with the matrix notation were defined previously.

An effective poststratification adjustment can reduce the bias in the resulting point estimates and will either reduce or minimally inflate the variance in comparison to the unadjusted weight. This effect is well known for traditional poststratification; we provide the comparative evaluation under an estimated-control setting in the next sections.



### 3. Bias in the ECPS of a population total

Traditional poststratification is known for reducing the bias associated with an incomplete sampling frame. This reduction is most successful when poststrata are formed such that the within-poststratum correlation of  $y_k$  with the probability of the  $k^{\text{th}}$  element being included on the sampling frame is very near zero (Kim, Li and Valliant 2007).

To evaluate the (unconditional) design-based bias for  $\hat{t}_{yP}$ , we must account for the random property of four components – the analytic and benchmark sample designs and the population coverage propensities for the corresponding sampling frames. Following the work of Kim, Li and Valliant (2007, equation 2), the approximate design bias of  $\hat{t}_{yP}$  as an estimator of the population total  $t_y = \sum_{k \in U} y_k$  is calculated as

$$\text{Bias}(\hat{t}_{yP}) = E(\hat{t}_{yP}) - t_y$$

$$\cong \sum_{g=1}^G t_{yg} \left\{ \frac{N_{Bg}}{N_g} - 1 \right\} + N_{Bg} \text{Cov}(y_g, \phi_{Ag}) \bar{\Phi}_{Ag}^{-1} \quad (4)$$

where  $N_g$  is the population size for the set of elements  $U_g$  within poststratum  $g$ ;  $N_{Bg} = E(\hat{N}_{Bg})$ , the expected value of the poststratum estimates under the benchmark survey design;  $\text{Cov}(y_g, \phi_{Ag}) = N_g^{-1} \sum_{k \in U_g} (y_k - \bar{y}_g)(\phi_{Ak} - \bar{\phi}_{Ag})$ , the population covariance between the outcome variable ( $y_k$ ) and the coverage propensities ( $\phi_{Ak}$ ) within poststratum  $g$ ;  $\bar{y}_g = t_{yg}/N_g$ , the  $g^{\text{th}}$  poststratum mean of  $y$ ;  $t_{yg} = \sum_{k \in U_g} y_k$ , the population total of  $y$  within poststratum  $g$ ; and  $\bar{\phi}_{Ag} = N_{Ag}/N_g$ , the average coverage propensity within the poststratum under the analytic survey design with  $N_{Ag} = E(\hat{N}_{Ag})$ . Note that the population total may also be expressed as  $t_y = \sum_g t_{yg}$ .

Components of the bias are zero only under certain conditions. (i) If  $N_{Bg} = N_g$  for all  $g$  (i.e., no coverage errors in the benchmark sampling frame), then the bias is dependent only on the association between the outcome variable and the coverage propensities,  $\text{Cov}(y_g, \phi_{Ag})$ . The value of  $\text{Bias}(\hat{t}_{yP})$  then reduces to the formula provided in Kim, Li and Valliant (2007, equation 2) for the traditional poststratified estimator,  $\hat{t}_{yPS}$ . (ii) If the coverage probabilities are constant within each poststratum (i.e.,  $\phi_{Ak} = \bar{\phi}_{Ag}$ ,  $k \in U_g$  for all  $g$ ), then the second bias component is zero. Only if *both* conditions are satisfied can we say that  $\hat{t}_{yP}$  is approximately unbiased. Some may argue that a “perfect” combination of poststrata could be formed such that the positive and negative components cancel; however, we believe this likelihood to be so rare as to be virtually impossible.

Having examined bias, we present an evaluation of the variance of  $\hat{t}_{yP}$ . For some estimators, the contribution of the bias (squared) to the total mean square error (MSE) is small relative to the variance.

### 4. Variance estimation for the ECPS

Variance estimators have been developed for traditional poststratification and are available in software designed to analyze survey data, e.g., R<sup>®</sup> (R Development Core Team 2009), SAS<sup>®</sup> (SAS Institute Inc. 2009), Stata<sup>®</sup> (StataCorp 2010), and SUDAAN<sup>®</sup> (Research Triangle Institute 2008). However, limited work has been completed on variance estimation for EC poststratification.

Four EC variance estimators for  $\hat{t}_{yP}$  that account for the variance in the control totals are presented in the following subsections after defining the population sampling variance. They include one newly developed linearization variance estimator, and three delete-one-PSU (delete-one) jackknife variance estimators. With the delete-one jackknife, replicates are created by sequentially deleting one PSU and adjusting the weights for the remaining PSUs within the corresponding design stratum. This results in a total of  $m_A = \sum_{h=1}^H m_{Ah}$  replicates calculated by summing the number of analytic-survey PSUs per stratum ( $m_{Ah}$ ) across the  $H$  strata ( $h = 1, \dots, H$ ).

An effective variance estimator will reproduce the corresponding population sampling variance in expectation. The approximate (or asymptotic) population sampling variance of  $\hat{t}_{yP} = \hat{\mathbf{N}}_B' \hat{\mathbf{Y}}_A$  has the following form:

$$\begin{aligned} AV(\hat{t}_{yP}) &= \mathbf{N}_B' \mathbf{V}_A \mathbf{N}_B + 2 \bar{\mathbf{Y}}_A' \text{Cov}(\hat{\mathbf{N}}_B, \hat{\mathbf{Y}}_A) \mathbf{N}_B + \bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A \\ &= \mathbf{N}_B' \mathbf{V}_A \mathbf{N}_B + \bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A \end{aligned} \quad (5)$$

where  $\mathbf{N}_B = E(\hat{\mathbf{N}}_B)$ , a vector of expected values for the benchmark poststratum counts within the  $G$  poststrata;  $\hat{\mathbf{N}}_B = (\hat{N}_{B1}, \dots, \hat{N}_{BG})'$  is a  $G$ -length vector of control totals estimated from the benchmark survey;  $\bar{\mathbf{Y}}_A$  is a  $G$ -length vector with population components of the form  $\bar{y}_{Ag} = t_{yg}/N_{Ag}$ ;  $\mathbf{V}_A$  is the population (variance-)covariance matrix of the estimated components of the vector  $\bar{\mathbf{Y}}_A$ ; and  $\mathbf{V}_B$  is the covariance matrix of the  $G$  benchmark control estimates  $\hat{\mathbf{N}}_B$ . The first component,  $\mathbf{N}_B' \mathbf{V}_A \mathbf{N}_B$ , is the approximate variance for the traditional poststratified estimator  $\hat{t}_{yPS}$ , i.e., the benchmark estimates are treated as fixed. The component,  $\bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A$ , is the variance associated with the benchmark estimates conditioned on the analytic survey sample; this is the EC poststratification variance component. Because we assume that the analytic and benchmark surveys are independent, the covariance of estimates from the two surveys is, by definition, zero. Hence, the component  $\text{Cov}(\hat{\mathbf{N}}_B, \hat{\mathbf{Y}}_A)$  above is eliminated from the expression.

Krewski and Rao (1981), Rao and Wu (1985), and others demonstrated the asymptotic consistency of the linearization and jackknife variance estimators for nonlinear functions. However, this examination needs to be extended to the EC poststratification. We discuss the set of EC variance

estimators for the population sampling variance below identified or developed for our research. The sample estimators were calculated by substituting sample estimates for the corresponding variance parameters. We begin with an evaluation of a traditional or naïve poststratified variance estimator that does not account for the variation in the estimated controls.

#### 4.1 A traditional variance estimator for EC poststratification (Naïve)

A variety of variance estimators have been developed for poststratification estimators. With all of the methods, the controls are assumed to be fixed and known without error. Therefore,  $\bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A$ , the second (positive) component in expression (5), is zero because  $\mathbf{V}_B = \mathbf{0}$  by assumption. The linearization variance estimator has the form

$$\text{var}_{\text{Naïve}}(\hat{t}_{yP}) = \hat{\mathbf{N}}_B' \hat{\mathbf{V}}_A \hat{\mathbf{N}}_B \quad (6)$$

where  $\hat{\mathbf{N}}_B$  is the vector of the  $G$  benchmark control total estimates, and  $\hat{\mathbf{V}}_A$  is the estimated covariance matrix of the estimates  $\hat{\mathbf{Y}}_A = (\hat{t}_{Ay1}/\hat{N}_{A1}, \dots, \hat{t}_{AyG}/\hat{N}_{AG})$ . Because the second component in the second line of (5) is not estimated, any variance formula developed for traditional poststratification will by definition underestimate the population sampling variance. However, highly precise benchmark estimates may contribute a negligible EC-poststratification variance component to the overall estimate. Thus, the difference between the estimates for traditional and EC poststratification will for these situations also be negligible.

#### 4.2 Taylor series linearization (ECTS)

A linearization variance estimator for the  $\hat{t}_{yP}$  has the form:

$$\text{var}_{\text{ECTS}}(\hat{t}_{yP}) = \hat{\mathbf{N}}_B' \hat{\mathbf{V}}_A \hat{\mathbf{N}}_B + \hat{\mathbf{Y}}_A' \hat{\mathbf{V}}_B \hat{\mathbf{Y}}_A \quad (7)$$

where  $\hat{\mathbf{V}}_B$  is the estimated benchmark covariance matrix for the set of  $G$  control totals. The remaining terms are defined for expression (6). The ECTS formula is a function of the variance under traditional poststratification and an additive inflation term associated with the variation in the benchmark controls, *i.e.*,  $\text{var}_{\text{ECTS}}(\hat{t}_{yP}) = \text{var}_{\text{Naïve}}(\hat{t}_{yP}) + \hat{\mathbf{Y}}_A' \hat{\mathbf{V}}_B \hat{\mathbf{Y}}_A$ .

Ideally, the benchmark survey analysis file would be available to calculate the values for  $\hat{\mathbf{V}}_B$ . However, researchers may have to rely on published estimates for only the marginal control totals, *i.e.*, point and variance estimates by one characteristic instead of the counts and covariance estimates for a set of characteristics. The implications of having limited information are discussed further in Section 4.4.

#### 4.3 Fuller two-phase jackknife method (ECF2)

Isaki, Tsay and Fuller (2004) applied a two-phase delete-one jackknife variance estimator developed by Fuller (1998) to an EC poststratification situation. The premise behind Fuller's methodology (ECF2) is to take a spectral (eigenvalue) decomposition of the benchmark covariance matrix ( $\hat{\mathbf{V}}_B$ ), develop benchmark adjustments that are a function of the resulting eigenvalues and eigenvectors, and add the adjustments to the vector of benchmark controls ( $\hat{\mathbf{N}}_B$ ) to create a set of replicate controls. A randomly chosen subset of the  $m_A$  replicates is poststratified to the  $G$  constructed replicate controls where the total number of PSUs must equal or exceed the number of poststrata, *i.e.*,  $m_A \geq G$ . Specifically, the benchmark control total for the  $r^{\text{th}}$  replicate is defined as

$$\hat{\mathbf{N}}_{B(r)} = \hat{\mathbf{N}}_B + c_h \hat{\mathbf{z}}'_{(r)} \quad (8)$$

where  $\hat{\mathbf{z}}'_{(r)} = \delta_{(r)} \sum_{g=1}^G \delta_{g|(r)} \hat{\mathbf{z}}'_g$ ;  $c_h = \sqrt{m_{Ah}/(m_{Ah} - 1)}$ , a constant related to the delete-one jackknife variance method;  $\delta_{(r)}$  is a zero/one indicator that identifies the  $G$  (out of  $m_A$ ) randomly chosen replicates to receive an adjustment;  $\delta_{g|(r)} = 1$  if the  $g^{\text{th}}$  component of the benchmark covariance decomposition is randomly chosen for the assignment given that replicate  $r$  is selected for adjustment; and  $\hat{\mathbf{z}}_g = \hat{\mathbf{q}}_g \sqrt{\hat{\lambda}_g}$ , a function of an eigenvector ( $\hat{\mathbf{q}}_g$ ) and the associated eigenvalue ( $\hat{\lambda}_g$ ) where  $\hat{\mathbf{V}}_B = \sum_{g=1}^G \hat{\mathbf{z}}_g \hat{\mathbf{z}}'_g$ , by definition. Thus, given that  $\delta_{(r)} = 1$  for a particular replicate, a single indicator  $\delta_{g|(r)}$  must also equal one; however, if  $\delta_{(r)} = 0$ , then *all* indicators  $\delta_{g|(r)}$  equal zero.

The delete-one jackknife can take multiple forms depending on the centering value. We chose the somewhat conservative variance estimator centered about the full-sample estimate for our research ( $v_4$  in Wolter 2007, section 4.5). The delete-one jackknife variance estimator,  $\text{var}_{\text{ECF2}}(\hat{t}_{yP})$ , is calculated as follows under the Fuller method for a stratified, multi-stage design.

$$\begin{aligned} \text{var}_{\text{ECF2}}(\hat{t}_{yP}) &= \sum_{h=1}^H \frac{(m_{Ah} - 1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\ddot{t}_{yP(r)} - \hat{t}_{yP})^2 \\ &= \sum_{h=1}^H \frac{(m_{Ah} - 1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yP(r)} - \hat{t}_{yP} + c_h \hat{\mathbf{z}}'_{(r)} \hat{\mathbf{B}}_{A(r)})^2 \quad (9) \end{aligned}$$

where the terms in (9) are defined below. Note that the association of the  $r^{\text{th}}$  replicate to a particular design stratum is defined through the stratum membership of the eliminated PSU. The replicate estimates in (9) are defined as  $\hat{t}_{Ay(r)} = \sum_h \sum_{i \in s_{Ah}} d_{i(r)} \sum_{k \in s_{Ahi}} \delta_{gk} d_k y_k$  and  $\hat{N}_{Ag(r)} = \sum_h \sum_{i \in s_{Ah}} d_{i(r)} \sum_{k \in s_{Ahi}} \delta_{gk} d_k$ , where the PSU-subsampling weights are calculated as



$$d_{i(r)} = \begin{cases} 0 & \text{if } r=i, i \in S_{Ah} \\ 1 & \text{if } h \neq h' \text{ for } r \in S_{Ah} \text{ and } i \in S_{Ah'} \\ m_{Ah}/(m_{Ah}-1) & \text{if } r \neq i \text{ but } h=h'. \end{cases} \quad (10)$$

The remaining terms in (9) are  $\hat{\mathbf{B}}_{A(r)} = \hat{t}_{Ag(r)}/\hat{N}_{Ag(r)}$ , the estimated mean of the outcome variable within poststratum  $g$  and replicate  $r$ ;

$$\ddot{i}_{yP(r)} = \sum_{g=1}^G \hat{N}_{Bg(r)} (\hat{t}_{Ag(r)}/\hat{N}_{Ag(r)}), \quad (11)$$

a function of replicate estimates with  $\hat{N}_{Bg(r)}$  defined as the  $g^{\text{th}}$  component in expression (8);  $\hat{t}_{yP(r)}$  is the replicate estimate under traditional poststratification, namely  $\sum_{g=1}^G \hat{N}_{Bg} (\hat{t}_{Ag(r)}/\hat{N}_{Ag(r)})$ ; and  $\hat{t}_{yP}$  is the estimated total given in expression (3) calculated from the complete sample file. Squaring the terms in (9) results in a variance component conditioned on the benchmark controls, a component due to the benchmark control variability, and a cross-term of lower order that is approximately equal to zero in expectation. The design-expectation of the resulting jackknife variance estimator is asymptotically equivalent to  $AV(\hat{t}_{yP})$  in (5) only if the respective components are calculated with values from design-consistent estimators. Fuller (1998) also demonstrated that the jackknife variance of the replicate controls,  $\text{var}_{\text{ECF2}}(\hat{\mathbf{N}}_B)$ , reproduces the estimated benchmark covariance matrix  $\hat{\mathbf{V}}_B$  for every sample.

Currently no software exists to calculate the ECF2. The six steps needed to calculate  $\text{var}_{\text{ECF2}}(\hat{t}_{yP})$  using any appropriate programmable package are as follows:

1. Calculate the full-sample estimate  $\hat{t}_{yP}$  using expression (3).
2. Determine the  $G$  eigenvalues  $\hat{\lambda}_g$  and eigenvectors  $\hat{\mathbf{q}}_g$  for  $\hat{\mathbf{V}}_B$ , and calculate the replicate adjustments  $\hat{\mathbf{z}}_g = \hat{\mathbf{q}}_g \sqrt{\hat{\lambda}_g}$ . Concatenate the  $G \times G$  matrix of  $\hat{\mathbf{z}}_g$ 's with a  $G \times (m_A - G)$  matrix of zeros, and randomly sort the columns. Call this new  $G \times m_A$  matrix  $\hat{\mathbf{Z}}$ .
3. Calculate a vector of length  $m_A$  with values equal to  $c_h = \sqrt{m_{Ah}/(m_{Ah}-1)}$  ordering from  $h=1$  to  $H$ . Populate each row of a  $G \times m_A$  matrix, called  $\mathbf{C}$ , with this vector, *i.e.*, the row values are repeated. The  $m_A$ -length vector of jackknife stratum weights,  $\mathbf{W}_R$ , is created with components equal to  $(m_{Ah}-1)/m_{Ah}$  where the deleted PSU is extracted from stratum  $h$ .
4. Calculate the Hadamard (or element-wise) product (Searle 1982, page 49) of  $\hat{\mathbf{Z}}$  and  $\mathbf{C}$  denoted as  $\hat{\mathbf{Z}} \bullet \mathbf{C}$ . Replicate the vector  $\hat{\mathbf{N}}_B$  into the columns of a  $G \times m_A$  matrix and add to  $\hat{\mathbf{Z}} \bullet \mathbf{C}$ . This new  $G \times m_A$  matrix, called  $\hat{\mathbf{N}}_{BR}$ , contains the replicate

benchmark controls discussed in expression (8) for all  $m_A$  replicates.

5. Calculate the replicate estimates  $\hat{y}_{Ag(r)} = \hat{t}_{Ag(r)}/\hat{N}_{Ag(r)}$  by removing in-turn one PSU from the analytic survey sample file, adjusting the weights for the remaining PSUs ( $\mathbf{W}_R$  values), and summing the weighted values for the numerator and denominator within poststratum  $g$ . Call the resulting  $G \times m_A$  matrix  $\hat{\mathbf{Y}}_R$ .
6. Calculate the  $m_A$  replicate estimates,  $\ddot{i}_{yP(r)}$ , by first multiplying the elements  $\hat{\mathbf{N}}_{BR}$  by  $\hat{\mathbf{Y}}_R$  and summing down the rows within a column. Next, subtract  $\hat{t}_{yP}$  from each of the  $m_A$  values and square the terms, multiply by the PSU-subsampling weight adjustments specified in (10), and sum across the  $m_A$  estimates. The resulting value is the estimated variance using the Fuller method,  $\text{var}_{\text{ECF2}}(\hat{t}_{yP})$ .

#### 4.4 Nadimpalli-Judkins-Chu jackknife method (ECNJC)

Nadimpalli *et al.* (2004) developed a delete-one jackknife variance estimator that randomly perturbs the control totals for the complete set of replicates instead of adjusting only a subsample of replicates as discussed for the ECF2. The benchmark survey replicate control totals have the following form:

$$\hat{\mathbf{N}}_{B(r)} = \hat{\mathbf{N}}_B + c_h R_h \hat{\mathbf{S}}_B \boldsymbol{\eta}_{(r)} \quad (12)$$

where  $c_h = \sqrt{m_{Ah}/(m_{Ah}-1)}$ , as with the ECF2;  $R_h = \sqrt{1/(H m_{Ah})}$ , a function of the total number of analytic-survey strata ( $H$ ) and PSUs ( $m_{Ah}$ );  $\hat{\mathbf{S}}_B$  is a diagonal matrix of estimated standard errors for the benchmark controls; and  $\boldsymbol{\eta}_{(r)}$  is a  $G$ -length vector of values randomly generated for each replicate from the standard normal distribution. The remaining terms are specified for the ECF2 following expression (8). Note that the covariance estimates included in the ECF2, *i.e.*, the off-diagonal values of  $\hat{\mathbf{V}}_B$ , are set to zero for the ECNJC.

The corresponding delete-one jackknife variance estimator of the poststratified total is calculated as follows:

$$\begin{aligned} \text{var}_{\text{ECNJC}}(\hat{t}_{yP}) &= \sum_{h=1}^H \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\ddot{i}_{yP(r)} - \hat{t}_{yP})^2 \\ &= \sum_{h=1}^H \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yP(r)} - \hat{t}_{yP} \\ &\quad + c_h R_h \boldsymbol{\eta}'_{(r)} \hat{\mathbf{S}}_B \hat{\mathbf{B}}_{A(r)})^2, \end{aligned} \quad (13)$$

where  $\ddot{i}_{yP(r)}$  is computed as described for the ECF2 in (11) but with  $\hat{N}_{Bg(r)}$  defined by the  $g^{\text{th}}$  component in (12). Unlike the ECF2, the sample variance of the ECNJC

replicate controls given in (12) reproduces the benchmark covariance matrix  $\mathbf{V}_B$  in expectation only if the covariance terms are truly zero (see Appendix A for details). If  $\mathbf{V}_B$  is not diagonal,  $\text{var}_{\text{ECNJC}}$  fails this test.

Use of the ECNJC would be plausible in two cases: (i) the complete benchmark covariance matrix for the controls is unavailable (e.g., estimates taken from a previous report), or (ii) the covariance terms are negative so that the resulting values defined by (12) would lead to conservative variance estimates. The diagonal matrix for  $\hat{\mathbf{S}}_B$  would be correct if the estimated poststratum counts were actually uncorrelated. However this is unlikely because of the multinomial structure of  $\hat{\mathbf{N}}_B$ . Given the setup for the ECNJC, the expectation of the variance estimator will *not* approximate  $\text{AV}(\hat{t}_{yP})$  in (5); the bias term is related to the difference between the design expectation of  $\hat{\mathbf{S}}_B^2$  and  $\mathbf{V}_B$ .

#### 4.5 Multivariate normal jackknife method (ECMV)

The multivariate normal method (ECMV) is a generalization of the ECNJC and to our knowledge is first discussed in this paper. The ECMV uses the complete covariance matrix  $\hat{\mathbf{V}}_B$  and relies on large-sample theory so that the control total adjustments may be modeled as coming from a  $G$ -dimensional multivariate normal (MVN) distribution. The replicate controls for the ECMV have the form

$$\hat{\mathbf{N}}_{B(r)} = \hat{\mathbf{N}}_B + c_h R_h \hat{\mathbf{e}}_{(r)} \quad (14)$$

where  $\hat{\mathbf{e}}_{(r)}$  is a  $G$ -length vector of random variables such that  $\hat{\mathbf{e}}_{(r)} \sim \text{MVN}_G(\mathbf{0}, \hat{\mathbf{V}}_B)$ ;  $c_h = \sqrt{m_{Ah}/(m_{Ah} - 1)}$ ; and  $R_h = \sqrt{1/(H m_{Ah})}$ .

The delete-one jackknife variance estimator for the ECMV is calculated as

$$\begin{aligned} \text{var}_{\text{ECMV}}(\hat{t}_{yP}) &= \sum_{h=1}^H \frac{(m_{Ah} - 1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\ddot{t}_{yP(r)} - \hat{t}_{yP})^2 \\ &= \sum_{h=1}^H \frac{(m_{Ah} - 1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yP(r)} - \hat{t}_{yP} \\ &\quad + c_h R_h \hat{\mathbf{e}}'_{(r)} \hat{\mathbf{B}}_{A(r)})^2, \end{aligned} \quad (15)$$

where  $\ddot{t}_{yP(r)}$  is computed as described for the ECF2 in (11) but with  $\hat{\mathbf{N}}_{B(r)}$  defined by the  $g^{\text{th}}$  component in (14). Unlike the Fuller method,  $\text{var}_{\text{ECMV}}(\hat{\mathbf{N}}_B) \neq \hat{\mathbf{V}}_B$ ; instead, the ECMV must rely on the design-based properties of the estimator. The design expectation of this estimator is evaluated with respect to the MVN distribution conditioned on the benchmark estimates ( $E_e$ ), and then with respect to the benchmark survey design ( $E_B$ ). As shown in Appendix B.1,

$$E_B[E_e(\text{var}_{\text{ECMV}}(\hat{\mathbf{N}}_B)|B)] = E_B(\hat{\mathbf{V}}_B). \quad (16)$$

If  $\hat{\mathbf{V}}_B$  is an approximately unbiased estimator of  $\mathbf{V}_B$ , then the population covariance matrix is reproduced with this method.

Under the Fuller two-phase method,  $\text{Var}[\text{var}_{\text{ECF2}}(\hat{\mathbf{N}}_B)] = \text{Var}(\hat{\mathbf{V}}_B)$  because  $\text{var}_{\text{ECF2}}(\hat{\mathbf{N}}_B) = \hat{\mathbf{V}}_B$ . To compare ECF2 and ECMV further, note that if we define  $y_k = 1$  in the analytic survey, then  $\hat{t}_{yP} = \mathbf{1}'\hat{\mathbf{N}}_B$ . As shown in Appendix B.2,

$$\begin{aligned} \text{Var}[\text{var}_{\text{ECMV}}(\mathbf{1}'\hat{\mathbf{N}}_B)] &= \\ \text{Var}_B[\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}] + \frac{2}{H\bar{m}_A^*} [E_B(\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1})^2] &> \text{Var}_B[\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}] \end{aligned} \quad (17)$$

where  $\bar{m}_A^*$  is the harmonic mean of the PSU sample sizes per stratum in the analytic survey. This suggests that the  $\text{var}_{\text{ECF2}}$  and the  $\text{var}_{\text{ECMV}}$  have similar large sample expectations, though in practice the ECMV is likely to be more variable than the ECF2. We examine this issue through a simulation study described in the next section.

## 5. Description of simulation study

We complement the theoretical evaluation of the five variance estimators discussed in the previous section with an analysis of simulation results.

### 5.1 Simulation parameters

The simulation population is a random subset of the 2003 National Health Interview Survey (NHIS) public-use file containing records for 21,664 adults. These records were divided into 25 strata, each containing six PSUs. Samples were selected from this “population” using a two-stage design. Two PSUs were selected *with replacement* using probabilities proportional to the total number of adults (PPS) within the PSU. From within each sample PSU, we selected simple random samples of ( $n_{Ahi} =$ ) 20 and 40 persons *without replacement* giving total sample sizes of 1,000 and 2,000, respectively. Two within-PSU sample sizes were considered for this study to evaluate the effects of smaller analytic survey variance components, calculated by increasing  $n_A$ , on the variance of  $\hat{t}_{yP}$ . For each combination of PSU and person-level samples (i.e., 50 PSUs and either 1,000 or 2,000 persons), we selected 4,000 simulation samples. We calculated the estimated population totals and associated variances for two binary NHIS variables: NOTCOV = 1 indicates that an adult *did not* have health insurance coverage in the 12 months prior to the NHIS interview (approximately 17 percent of the population); and PDMED12M = 1 indicates that an adult *delayed* medical care because of cost in the 12 months prior to the interview (approximately 7 percent of the population).



We exclude nonresponse from consideration in our current simulation study to minimize factors that might affect our comparisons. (Note: The interview questions for these variables can be found in the family core instrument at [ftp://ftp.cdc.gov/pub/Health\\_Statistics/NCHS/Survey\\_Questnaires/NHIS/2003/qfamilyx.pdf](ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Survey_Questnaires/NHIS/2003/qfamilyx.pdf). Responses from questions FHI.070 and FAU.010/FAU.020 were used to generate the variables NOTCOV and PDMED12M, respectively).

Poststratification may reduce variances slightly. However, in household surveys, this technique is mainly used to correct for sampling frame undercoverage, as well as other problems inherent with surveys. Each of the 4,000 simulation samples was selected to mimic a sampling frame for the analytic survey that suffers from differential undercoverage, such as those used for many telephone surveys. Sixteen ( $G = 16$ ) poststratification cells were defined by an eight-level age variable crossed with gender. The coverage rates for the 16 cells were created based on the population means for each age group by gender and range in value from 0.5 to 0.9. A coverage rate equal to 1.0 would indicate full coverage. Before each sample was selected, the frame was designated as a stratified random subsample of the full population of 21,664. For example, 90 percent of the male population 65-69 years of age was randomly selected to be in the sampling frame for the NOTCOV simulations. This process of subsetting the population to the frame was independently implemented for each sample and for each outcome variable.

We suspect that the decision for researchers to use either a traditional or an EC poststratification variance estimator depends on the precision of the control totals. We calculated the benchmark covariance matrix ( $\hat{V}_B$ ) from the complete NHIS public-use data file (92,148 records) and ratio adjusted the values to reflect a sample size comparable with our simulation population ( $N = 21,664$ ). The off-diagonal values of  $\hat{V}_B$  range from -0.05 to 0.75 with a mean value of 0.22. From this matrix we calculated four covariance matrices for the simulation by dividing the original matrix by the adjustment factors 1.0, 3.6, 18, and 72. The adjustments reflect benchmark surveys with an approximate effective sample size of 21,700, 6,000 ( $\approx 21,700/3.6$ ), 1,200, and less than 500, respectively.

The simulation was conducted in R<sup>®</sup> (Lumley 2009; R Development Core Team 2009) because of its extensive capabilities for analyzing survey data and efficiency with simulated analyses. Code was developed to calculate the linearization and replicate variance estimates for the EC poststratified estimator discussed above because the relevant code does not currently exist.

## 5.2 Evaluation criteria

The empirical results for the five variance estimators discussed in the previous section (Naïve, ECTS, ECF2, ECNJC, and ECMV) are compared using three measures across the  $j = 1, \dots, 4,000$  simulation samples, and the two outcome variables (NOTCOV and PDMED12M). The measures include: (i) the estimated percent relative bias of the variance estimator,  $(1/4,000 \sum_j \text{var}(\hat{t}_{yP_j}) - \text{mse})/\text{mse}$  where  $\text{var}(\hat{t}_{yP_j})$  is one of the five variance estimates evaluated for sample  $j$  and  $\text{mse}$  is the mean square error of  $\hat{t}_{yP}$  defined below; (ii) the 95% confidence interval coverage rate,  $1/4,000 \sum_j I(|\hat{z}_j| \leq z_{1-\alpha/2})$  where  $\hat{z}_j = (\hat{t}_{yP_j} - t_y)/\sqrt{\text{var}(\hat{t}_{yP_j})}$ ; and, (iii) the standard deviation of the estimated standard errors, calculated as the square root of  $1/(4,000 - 1) \sum_j (\sqrt{\text{var}(\hat{t}_{yP_j})} - 1/4,000 \sum_j \sqrt{\text{var}(\hat{t}_{yP_j})})^2$ . The relative bias and the root mean square error of our point estimators are calculated as  $1/4,000 \sum_j (\hat{t}_{yP_j} - t_y)/t_y$  and  $\sqrt{\text{mse}} = \sqrt{1/4,000 \sum_s (\hat{t}_{yP_j} - t_y)^2}$ , respectively.

## 6. Simulation study results

### 6.1 Point estimator

To justify the need for poststratification, we initially evaluated the Horvitz-Thompson estimate ( $\sum_{s,j} d_{kj} y_k$ ) for the two outcome variables. This estimator is known to be design-unbiased under pristine conditions. The percent relative bias indicates that the HT estimator is negatively biased, underestimating the population total by 38 percent for NOTCOV and 41 percent for PDMED12M. These large values show that some correction is needed to adjust for the non-negligible levels of bias. The percent relative bias for the poststratified estimator  $\hat{t}_{yP}$  was much lower – the  $\hat{t}_{yP}$  is positively biased by no more than two percent for both outcome variables.

### 6.2 Variance estimators

Adding to the theoretical evaluation discussed in Section 4, the empirical results for an effective variance estimator should possess a *percent relative bias* either near zero or somewhat positive for a conservative measure (see Section 5.2 for the formula of the percent relative bias).

The percent relative biases generated from our simulation study are provided in Table 1. Bias estimates for the Naïve and ECNJC variance estimators are larger than for the other EC estimators for all our simulations. Estimates for the ECTS are somewhat smaller than the values calculated for the ECF2 and ECMV estimators for relatively small benchmark surveys. However, the differences are negligible as the size of the benchmark survey increases.

**Table 1**  
Percent relative bias estimates for five variance estimators by outcome variable and relative size of the benchmark survey to the analytic survey

Outcome Variable	Variance Estimator	Relative Size ( $n_A = 1,000$ )				Relative Size ( $n_A = 2,000$ )			
		0.3	1.2	6.0	21.7	0.2	0.6	3.0	10.8
NOTCOV	Naïve	-50.3	-23	-10.7	-9.2	-56.0	-31	-14.2	-12.2
	ECTS	-4.5	-4.5	-6.1	-7.7	-0.2	-8.4	-8.2	-10.1
	ECF2	-4.7	-4.6	-5.8	-7.5	0.1	-8.2	-8.3	-10.1
	ECNJC	-36.7	-17.1	-8.9	-8.2	-40	-24.2	-11.9	-11.1
	ECMV	-4.3	-4.1	-6.0	-7.5	-0.2	-8.1	-8.1	-10.0
PDMED12M	Naïve	-34.4	-14.5	-5.7	-3.9	-48.1	-23.4	-10	-10.1
	ECTS	-3.3	-3.7	-2.7	-2.6	-4.7	-6.4	-5.1	-7.8
	ECF2	-3.5	-3.5	-2.4	-2.3	-4.6	-6.8	-5.2	-7.8
	ECNJC	-24.5	-10.5	-4.0	-2.7	-35.1	-17.6	-7.6	-8.4
	ECMV	-3.0	-3.3	-2.4	-2.2	-4.3	-6.3	-5.0	-7.7

The traditional poststratified estimator (Naïve) was most negatively biased among those compared as expected. When the benchmark survey is smaller than the analytic survey (and therefore produces estimates less precise than the analytic survey), the Naïve estimator is negatively biased by as much as 56 percent. The level of bias improved as the relative size of the benchmark survey increased; however, the Naïve estimator still resulted in, at best, a four percent underestimate. The ECNJC estimator fared slightly better than the Naïve estimator though the bias (-2.7 to -40 percent) is still larger than the other EC variance estimators, which range between -10.1 and 0.1 percent.

For a small benchmark survey relative to the size of the analytic survey (*i.e.*, relative size less than one), the levels of (absolute) bias dramatically increased for the Naïve and ECNJC estimators. The opposite effect is noted for the other EC variance estimators. The variance component associated with the benchmark survey, *e.g.*,  $\hat{\mathbf{Y}}_A' \hat{\mathbf{V}}_B \hat{\mathbf{Y}}_A$  shown for  $\text{var}_{\text{ECTS}}$  in (7), becomes the dominate term within the EC variance estimators as the precision of the benchmark survey estimates decreases. Thus the benchmark variance component somewhat corrects for the underestimation associated with the analytic variance component. Additional research is needed to determine if a threshold exists for when such a counterbalance of bias can occur. The overall negative bias of our estimates is similar to the bias of linearization variance estimators as shown in another context by Rao and Wu (1985, section 4) and Wu (1985). However, further research is also needed to determine how to minimize the underestimation.

Note that the relative sizes of 21.7 when  $n_A = 1,000$  and 10.8 when  $n_A = 2,000$  both imply benchmark survey sample sizes of about 21,600. Thus the  $O(M^2/m_B)$  component of the variance,  $\bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A$ , is more prominent for the estimates in Table 1 based on  $n_A = 2,000$ . This leads to larger relative biases in these estimates, relative to those produced under  $n_A = 1,000$ , even though the analytic survey sample size is larger.

The patterns exhibited for the percent relative bias are reflected in the *coverage rates for the 95 percent confidence intervals* for the estimated totals but are not provided for sake of brevity. The Naïve and ECNJC estimators are more likely to experience confidence intervals coverage rates below 95 percent. These rates approach the appropriate level as the precision of the benchmark survey estimates improves. However, the remaining EC variance estimators had coverage rates near acceptable levels regardless of the relative size of the surveys and therefore are more robust.

The discussion so far suggests that there are minimal theoretical, as well as empirical, differences between the ECTS, ECF2, and ECMV methods. We finally look to the *standard deviation of the estimated standard errors* (SEs) in an attempt to distinguish the estimators. An examination of this variability can provide insight on the (empirical) stability of the variance estimators, *i.e.*, an unstable variance estimator could generate a poor variance estimate based on the nuances of a particular sample. Table 2 contains the percent relative increase in the standard deviations for the ECF2 and the ECMV both in comparison to the ECTS.

The variation in the ECMV variance estimates was noticeably larger than for ECF2 but only for relatively small benchmark surveys. The difference increased as the size of the analytic survey increased. This suggests that the ECF2 may be preferred over the ECMV due to increased stability in the variance estimates. However, further research is being conducted on the threshold for when the instability can affect the estimates.

## 7. Conclusions and future work

The theoretical and analytical work discussed in this paper support the need for a new methodology to address post-stratification using estimated control totals, *i.e.*, estimated-control (EC) poststratification. Traditional variance estimators can severely underestimate the population sampling variance resulting in, for example, incorrect decisions for hypothesis tests and sub-optimal sample allocations when the design is implemented in the future.



Table 2

Percent increase in instability of variance estimates relative to the ECTS by outcome variable and relative size of the benchmark survey

Outcome Variable	Variance Estimator	Relative Size ( $n_A = 1,000$ )				Relative Size ( $n_A = 2,000$ )			
		0.3	1.2	6.0	21.7	0.2	0.6	3.0	10.8
NOTCOV	ECF2	12.0	5.5	2.3	0.2	15.1	8.4	2.1	0.6
	ECMV	21.2	7.4	1.8	0.3	30.8	8.5	2.4	0.7
PDMED12M	ECF2	7.7	3.8	1.1	0.4	12.0	6.3	2.1	0.7
	ECMV	11.5	4.0	0.9	0.5	22.6	7.6	2.2	1.1

The EC linearization variance estimator  $\text{var}_{\text{ECTS}}$  in expression (7) shows promise for EC poststratification. This estimator is especially effective at reducing the percent relative bias experienced with the Naïve variance estimator in (6) when the benchmark survey is small relative to the analytic survey. The replication variance estimator  $\text{var}_{\text{ECF2}}$  given in (9) is recommended specifically for studies requiring replicate weights such as when public-use analysis files are released without sampling design information to further protect data confidentiality and respondent privacy. The alternative replication estimator  $\text{var}_{\text{ECMV}}$  also performed well and is somewhat easier to implement than  $\text{var}_{\text{ECF2}}$ .

Implementation of the recommended variance estimators requires specialized computer programs because the capabilities are currently not available in standard software. The linearization estimator may be more approachable because implementation involves a modification to available variance estimates, e.g.,  $\text{var}_{\text{ECTS}}(\hat{t}_{y\text{ECPS}}) = \text{var}_{\text{Naïve}}(\hat{t}_{y\text{ECPS}}) + \hat{\mathbf{Y}}_A' \hat{\mathbf{V}}_B \hat{\mathbf{Y}}_A$ . We provide a step-by-step discussion of the procedures required for the  $\text{var}_{\text{ECF2}}$  (see Section 4.3) to facilitate the creation of the computer program.

Extensions to this research to be presented at a later date include a generalization to linear calibration, to other statistics including a ratio-estimated mean, and to domain estimation. We additionally are investigating whether threshold values are identifiable which determine (i) when there are negligible differences between traditional and EC variance estimation, and (ii) when the benchmark controls are too imprecise to use for calibration. We also plan to investigate the theoretical implications of measurement errors in the analytic as well as the benchmark surveys.

### Acknowledgements

This work was completed as part of the first author's doctoral dissertation at the Joint Program in Survey Methodology, University of Maryland. She thanks the members of her committee, Richard Valliant, Phillip Kott, Frauke Kreuter, Stephen Miller and Paul Smith for their guidance. The authors also thank the associate editor and referees for their constructive comments which clarified the presentation.

## Appendix A

### Derivation of $\text{var}_{\text{ECNJC}}(\hat{\mathbf{N}}_B)$

For the following derivations, let  $E_e$  represent the expectation with respect to a standard normal distribution. All other terms are defined in the body of the paper.

$$\begin{aligned} \text{var}_{\text{ECNJC}}(\hat{\mathbf{N}}_B) &= \sum_{h=1}^H \frac{m_{Ah} - 1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)(\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)' \\ &= \frac{1}{H} \hat{\mathbf{S}}_B \left( \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} \mathbf{K}_{(r)} \right) \hat{\mathbf{S}}_B \end{aligned}$$

where  $\mathbf{K}_{(r)} = \boldsymbol{\eta}_{(r)} \boldsymbol{\eta}_{(r)}'$ , a  $G \times G$  cross-product matrix of standard normal values; and  $\hat{\mathbf{S}}_B^2 = \text{diag}(\hat{\mathbf{V}}_B)$ . Because  $E_e(\mathbf{K}_{(r)}) = \mathbf{I}_G$ , a  $G$ -dimension identity matrix, we have  $E_e[\text{var}_{\text{ECNJC}}(\hat{\mathbf{N}}_B)] = \text{diag}(\hat{\mathbf{V}}_B)$ . Therefore,  $\text{var}_{\text{ECNJC}}(\hat{\mathbf{N}}_B)$  does not reproduce  $\hat{\mathbf{V}}_B$  in expectation.

## Appendix B

### Evaluation of the ECMV

For the following derivations, let  $E_B$  and  $\text{Var}_B$  represent the expectation and variance with respect to the benchmark survey sampling design. Also, let  $E_e$  and  $\text{Var}_e$  represent the expectation and variance with respect to the  $G$ -dimensional multivariate normal distribution,  $\text{MVN}_G(\mathbf{0}, \hat{\mathbf{V}}_B)$ . All other terms are defined in the body of the paper.

#### B.1: Derivation of $E[\text{var}_{\text{ECMV}}(\hat{\mathbf{N}}_B)]$ given in (15)

Using expression (14) and  $c_h^2 = m_{Ah}/(m_{Ah} - 1)$ ,

$$\begin{aligned} E[\text{var}_{\text{ECMV}}(\hat{\mathbf{N}}_B)] &= E_B \left[ E_e \left( \sum_{h=1}^H \frac{(m_{Ah} - 1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)(\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)' \middle| B \right) \right] \\ &= \frac{1}{H} E_B \left[ \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_e(\hat{\mathbf{e}}_{(r)} \hat{\mathbf{e}}_{(r)}' | B) \right] \\ &= \frac{1}{H} \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_B(\hat{\mathbf{V}}_B) = E_B(\hat{\mathbf{V}}_B). \end{aligned}$$

## B.2: Derivation of $\text{Var}[\text{var}_{\text{ECMV}}(\hat{\mathbf{N}}_B)]$ given in (15)

When  $y_k = 1$  so that  $\hat{t}_{yP} = \mathbf{1}'\hat{\mathbf{N}}_B$ ,  $\text{var}_{\text{ECMV}}(\mathbf{1}'\hat{\mathbf{N}}_B) = H^{-1} \sum_{h=1}^H m_{Ah}^{-1} \sum_{r=1}^{m_{Ah}} \mathbf{1}'\hat{\mathbf{e}}_{(r)}\hat{\mathbf{e}}_{(r)}'\mathbf{1}$ . Using the formula for the variance of a quadratic form (Searle 1982, section 13.5), we have

$$\begin{aligned} \text{Var}[\text{var}_{\text{ECMV}}(\mathbf{1}'\hat{\mathbf{N}}_B)] &= \text{Var}_B \left[ \frac{1}{H} \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_{\epsilon}(\mathbf{1}'\hat{\mathbf{e}}_{(r)}\hat{\mathbf{e}}_{(r)}'\mathbf{1}|B) \right] \\ &\quad + E_B \left[ \frac{1}{H^2} \sum_{h=1}^H \frac{1}{m_{Ah}^2} \sum_{r=1}^{m_{Ah}} \text{Var}_{\epsilon}(\mathbf{1}'\hat{\mathbf{e}}_{(r)}\hat{\mathbf{e}}_{(r)}'\mathbf{1}|B) \right] \\ &= \text{Var}_B \left[ \frac{1}{H} \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} \mathbf{1}'\hat{\mathbf{V}}_B \mathbf{1} \right] \\ &\quad + E_B \left[ \frac{1}{H^2} \sum_{h=1}^H \frac{1}{m_{Ah}} \{ 2\text{tr}(\mathbf{1}'\hat{\mathbf{V}}_B \mathbf{1}\mathbf{1}'\hat{\mathbf{V}}_B) \} \right] \\ &= \text{Var}_B[\mathbf{1}'\hat{\mathbf{V}}_B \mathbf{1}] + \frac{2}{H\bar{m}_A^*} [E_B(\mathbf{1}'\hat{\mathbf{V}}_B \mathbf{1})^2], \end{aligned}$$

where  $\bar{m}_A^* = (H^{-1} \sum_{h=1}^H m_{Ah}^{-1})^{-1}$  is the harmonic mean of  $m_{Ah}$ .

## References

- Binder, D.A. (1995). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22, 1, 17-22.
- Bray, R., Hourani, L., Rae, K., Dever, J., Brown, J., Vincus, A., Pemberton, M., Marsden, M., Faulkner, D. and Vandermaas-Peeler, R. (2003). 2002 Department of Defense Survey of Health Related Behaviors Among Military Personnel. Tech. Rep. RTI/7841/006-FR, U.S. Department of Defense prepared by RTI International. URL <http://dodwws.rti.org/2002WWFinalReportComplete05-04.pdf>.
- Canty, A.J., and Davison, A.C. (1999). Resampling-based variance estimation for Labour Force Surveys. *The Statistician*, 48, 379-391.
- Centers for Disease Control and Prevention (2006). Technical Information and Data for the Behavioral Risk Factor Surveillance System (BRFSS) – BRFSS Weighting Formula. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, September 11, 2006.
- Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 1, 17-26.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.
- Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88(423), 1013-1020.
- Estevao, V.M., and Särndal, C.-E. (2000). A Functional form approach to calibration. *Journal of Official Statistics*, 16(4), 379-399.
- Fuller, W.A. (1998). Replication variance estimation for the two-phase samples. *Statistica Sinica*, 8, 1153-1164.
- Hidiroglou, M.A., and Patak, Z. (2006). Raking ratio estimation: An application to the Canadian Retail Trade Survey. *Journal of Official Statistics*, 22(1), 71-80.
- Isaki, C.T., Tsay, J.H. and Fuller, W.A. (2004). Weighting sample data subject to independent controls. *Survey Methodology*, 30, 1, 35-44.
- Jayasuriya, B.R., and Valliant, R. (1996). An application of regression and calibration estimation to post-stratification in a Household Survey. *Survey Methodology*, 22, 2, 127-137.
- Keeter, S., Dimock, M. and Christian, L. (2008). Calling Cell Phones in '08 Pre-Election Polls. NEWS Release (December 18, 2008): Pew Research Center for the People & the Press. URL <http://people-press.org/reports/pdf/cell-phone-commentary.pdf>.
- Killion, R.A. (2006). Weighting Specifications for The American Time Use Survey (ATUS) for 2006. U.S. Bureau of the Census, Internal Memo (Doc.#ATUS-16).
- Kim, J.J., Li, J. and Valliant R. (2007). Cell collapsing in poststratification. *Survey Methodology*, 33, 2, 139-150.
- Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 2, 133-142.
- Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9(5), 1010-1019.
- Krotki, K. (2007). Combining RDD and Web Panel Surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association (in print).
- Lessler, J.T., and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. New York: John Wiley & Sons, Inc.
- Lumley, T. (2009). Survey: Analysis of complex survey samples. R package version 3.19. University of Washington: Seattle.
- Mirza, H., and Hörngren, J. (2002). The Sampling and the Estimation Procedure in the Swedish Labour Force Survey. Technical report, Statistics Sweden, Stockholm: Sweden.
- Nadimpalli, V., Judkins, D. and Chu, A. (2004). Survey Calibration to CPS Household Statistics. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 4090-4094.
- R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available: <http://www.R-project.org>.
- Rao, J.N.K., and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86(2), 403-415.
- Rao, J.N.K., and Wu, C.F.J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80(391), 620-630.



- Research Triangle Institute (2008). *SUDAAN Language Manual*. Release 10.0, Research Triangle Park, NC: Research Triangle Institute.
- Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. England: John Wiley & Sons, Inc.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3), 527-537.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag, Inc.
- SAS Institute Inc. (2009). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.
- Searle, S.R. (1982). *Matrix Algebra Useful for Statistics*. New York: John Wiley & Sons, Inc.
- StataCorp (2010). *Stata Statistical Software: Release 11*. Survey Data, College Station, TX: StataCorp LP.
- Stukel, D.M., Hidioglou, M.A. and Särndal, C.-E. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 2, 117-125.
- Taylor, M.F., Brice, J., Buck, N. and Prentice-Lane, E. (2007). *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. University of Essex, Colchester.
- Terhanian, G., Bremer, J., Smith, R. and Thomas, R. (2000). *Correcting Data from Online Survey for the Effects of Nonrandom Selection and Nonrandom Assignment*. Research Paper: Harris Interactive.
- Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94(446), 635-644.
- Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.
- Wolter, K.M. (2007). *Introduction to Variance Estimation*. New York: Springer Science+Business Media, LLC.
- Wu, C.F.J. (1985). Variance estimation for the combined ratio and combined regression estimators. *Journal of the Royal Statistical Society, Series B*, 47(1), 147-154.
- Yung, W., and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.
- Yung, W., and Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95(451), 903-915.

# Some contributions to jackknifing two-phase sampling estimators

Patrick J. Farrell and Sarjinder Singh<sup>1</sup>

## Abstract

In this paper, the problem of estimating the variance of various estimators of the population mean in two-phase sampling has been considered by jackknifing the two-phase calibrated weights of Hidirolou and Särndal (1995, 1998). Several estimators of population mean available in the literature are shown to be the special cases of the technique developed here, including those suggested by Rao and Sitter (1995) and Sitter (1997). By following Raj (1965) and Srivenkataramana and Tracy (1989), some new estimators of the population mean are introduced and their variances are estimated through the proposed jackknife procedure. The variance of the chain ratio and regression type estimators due to Chand (1975) are also estimated using the jackknife. A simulation study is conducted to assess the efficiency of the proposed jackknife estimators relative to the usual estimators of variance.

Key Words: Auxiliary information; Calibration; Estimation of mean and variance; Jackknife; Two-phase sampling.

## 1. Introduction

Hidirolou and Särndal (1995, 1998) have pointed out that two-phase sampling for the estimation of finite population attributes is a powerful and cost-effective technique, and hence plays an eminent role in survey sampling. Two-phase sampling can be described as follows. Consider a finite population that we shall denote by  $\Omega = \{1, 2, \dots, i, \dots, N\}$ . Suppose that information is available on a variable  $Z$  across the entire population; that is, the values  $Z_i$  for all  $i = 1, \dots, N$ , are known, implying that the population mean,  $\bar{Z}$ , is also known. A first-phase probability sample  $s_1, s_1 \subset \Omega$ , of size  $m$  is drawn from the population with selection probabilities  $\pi_{1i}$ . Thus, the first-phase sampling weights can be defined as  $d_{1i} = 1/\pi_{1i}$ . Assume that for this sample, information is collected on a variable  $X$ , which is then paired with the information on  $Z$  for each of the  $m$  units, giving rise to the data  $\{(x_i, z_i) | i \in s_1\}$  for  $i = 1, \dots, m$ . Once the first-phase sample  $s_1$  has been drawn, a second-phase sample  $s_2, s_2 \subset s_1 \subset \Omega$ , of size  $n$  is selected from  $s_1$  with selection probabilities  $\pi_{2i} = \pi_{i|s_1}$ , allowing for the second-phase sampling weights to be defined as  $d_{2i} = 1/\pi_{2i}$ . In the second-phase sample, information is now collected on a variable  $Y$  for each selected unit. This information is linked to that previously available on  $Z$  and  $X$  for these units, giving rise to the data  $\{(x_i, y_i, z_i) | i \in s_2\}$  for  $i = 1, \dots, n$ . Suppose that interest lies in estimating the population mean  $\bar{Y}$ , and on the variance of the estimator employed.

Let  $w_{1i}^o = d_{1i}/\sum_{i \in s_1} d_{1i}$  denote the first-phase normalized original design weights. The usual estimator of the population mean  $\bar{X}$  is given by

$$\hat{\bar{X}}_1^o = \sum_{i \in s_1} w_{1i}^o x_i,$$

while a calibrated first-phase estimator of  $\bar{X}$  is

$$\hat{\bar{X}}_1^c = \sum_{i \in s_1} w_{1i}^c x_i,$$

where the  $w_{1i}^c$  are calibrated weights such that the chi-square distance function

$$D_1 = \sum_{i \in s_1} \{(w_{1i}^c - w_{1i}^o)^2 / (w_{1i}^o q_{1i})\}, \quad (1.1)$$

is minimized subject to

$$\sum_{i \in s_1} w_{1i}^c z_i = \bar{Z}. \quad (1.2)$$

In (1.1), the  $q_{1i}$  are a set of suitably chosen weights. Minimization of (1.1) subject to (1.2) leads to the first-phase calibrated weights

$$w_{1i}^c = w_{1i}^o + \left\{ (q_{1i} w_{1i}^o z_i) / \left( \sum_{i \in s_1} q_{1i} w_{1i}^o z_i^2 \right) \right\} \left( \bar{Z} - \sum_{i \in s_1} w_{1i}^o z_i \right).$$

Thus, a first-phase calibrated estimator of  $\bar{X}$  is given by

$$\hat{\bar{X}}_1^c = \sum_{i \in s_1} w_{1i}^o x_i + \hat{\beta}_1 \left( \bar{Z} - \sum_{i \in s_1} w_{1i}^o z_i \right),$$

where

$$\hat{\beta}_1 = \left( \sum_{i \in s_1} q_{1i} w_{1i}^o x_i z_i \right) / \left( \sum_{i \in s_1} q_{1i} w_{1i}^o z_i^2 \right).$$

Now, let  $w_{2i}^o = d_{1i} d_{2i} / \sum_{i \in s_2} d_{1i} d_{2i}$  denote the second-phase normalized design weights. The usual estimator of  $\bar{Y}$  is given by

1. Patrick J. Farrell, School of Mathematics and Statistics, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, Canada, K1S 5B6. E-mail: pfarrell@math.carleton.ca; Sarjinder Singh, Department of Mathematics, Texas A&M University - Kingsville, Kingsville, Texas, U.S.A., 78363. E-mail: sarjinder@yahoo.com.



$$\hat{Y}_2'' = \sum_{i \in s_2} w_{2i}^o y_i.$$

Let us consider the second-phase calibrated estimator of  $\bar{Y}$  as

$$\hat{Y}^c = \sum_{i \in s_2} w_{2i}^c y_i, \quad (1.3)$$

where the  $w_{2i}^c$  are the second-phase calibrated weights such that the chi-square distance function

$$D_2 = \sum_{i \in s_2} \{(w_{2i}^c - w_{2i}^o)^2 / (w_{2i}^o q_{2i})\}, \quad (1.4)$$

is minimized subject to the calibration constraint

$$\sum_{i \in s_2} w_{2i}^c x_i = \hat{X}_1^c. \quad (1.5)$$

Minimization of (1.4) subject to (1.5) leads to the second-phase calibrated weights

$$w_{2i}^c = w_{2i}^o + \left\{ (q_{2i} w_{2i}^o x_i) / \left( \sum_{i \in s_2} q_{2i} w_{2i}^o x_i^2 \right) \right\} \left( \hat{X}_1^c - \sum_{i \in s_2} w_{2i}^o x_i \right).$$

Thus, the second-phase calibrated estimator of  $\bar{Y}$  specified in (1.3) can be written as

$$\hat{Y}^c = \hat{Y}_2'' + \hat{\beta}_2 (\hat{X}_1'' - \hat{X}_1'') + \hat{\beta}_1 \hat{\beta}_2 (\bar{Z} - \hat{Z}_1''), \quad (1.6)$$

where  $\hat{Z}_1'' = \sum_{i \in s_1} w_{1i}^o z_i$ ,  $\hat{X}_1'' = \sum_{i \in s_1} w_{1i}^o x_i$ ,  $\hat{X}_2'' = \sum_{i \in s_2} w_{2i}^o x_i$ ,  $\hat{Y}_2'' = \sum_{i \in s_2} w_{2i}^o y_i$ , and

$$\hat{\beta}_2 = \left( \sum_{i \in s_2} q_{2i} w_{2i}^o x_i y_i \right) / \left( \sum_{i \in s_2} q_{2i} w_{2i}^o x_i^2 \right).$$

Hidiroglou and Särndal (1995, 1998) and Singh (2000) have considered the problem of estimating the variance of the calibrated estimator  $\hat{Y}^c$  in (1.6) by using a design-based approach. In a more general context, Rao and Sitter (1995) and Sitter (1997) have pointed out that under simple random sampling without replacement (SRSWOR), a jackknife technique can be used to estimate the variances of the ratio and regression estimators for a population mean. These authors have also reported that the use of the jackknife for estimating variance is more convenient and efficient than the traditional techniques based on estimates of moments.

Of late, a number of authors have investigated the use of jackknife procedures for estimating variances (See Arnab and Singh 2006, Berger 2007, Berger and Skinner 2005, Chen and Shao 2001, and Kovar and Chen 1994). Fuller (1998), Kim, Navarro and Fuller (2000, 2006), Kim and Sitter (2003), and Kott and Stukel (1997) have suggested an approach for estimating the variance in two-stage sampling.

Fuller (1998) and Kim and Sitter (2003) address the regression estimator. In particular, consider the generalized regression estimator of population total

$$\hat{Y}_{DS} = \sum_{i \in s_2} \alpha_i y_i,$$

due to Deville and Särndal (1992). Following Kim *et al.* (2000, 2006), for each  $k \in s_2$ , specify the jackknife estimator of population total as

$$\hat{Y}_{Kim} = \sum_{i \in s_2 \setminus k} \alpha_i^{(k)} y_i, \quad (1.7)$$

and the chi-square distance between the design and calibration weights as

$$D_{(k)} = (1/2) \sum_{i \in s_2 \setminus k} \{(\alpha_i^{(k)} - w_i^{(k)} w_i^{*(k)})^2 / (w_i^{(k)} q_i^{(k)})\}. \quad (1.8)$$

Minimizing (1.8) subject to the condition

$$\sum_{i \in s_2 \setminus k} \alpha_i^{(k)} x_i = \sum_{i \in s_1 \setminus k} w_i^{(k)} x_i,$$

leads to jackknifed calibrated weights given by

$$\alpha_i^{(k)} = w_i^{(k)} w_i^{*(k)} + \left\{ (w_i^{(k)} q_i^{(k)} x_i) / \left( \sum_{i \in s_2 \setminus k} w_i^{(k)} q_i^{(k)} \right) \right\} \left\{ \sum_{i \in s_2 \setminus k} w_i^{(k)} x_i - \sum_{i \in s_2 \setminus k} w_i^{(k)} w_i^{*(k)} x_i \right\}.$$

It would appear that Kim *et al.* (2006) readjusted these weights as

$$\alpha_i^{(k)} = \begin{cases} \alpha_i^{(k)} & \text{if } k \in s_2 \\ w_i^{(k)} & \text{if } j \in (s_1 - s_2). \end{cases}$$

For such a readjustment, the estimator in (1.7) is equivalent to that of Rao and Sitter (1995).

In the present paper, we consider a new jackknife technique to estimate the variance of the estimator  $\hat{Y}^c$  under the two-phase setup by following Hidiroglou and Särndal (1995, 1998). Similar to Kim *et al.* (2006), the estimator proposed by Rao and Sitter (1995) is shown to be a special case of the proposed method. However, our approach differs from that of Fuller (1998) Kim and Sitter (2003), Kim *et al.* (2000, 2006) in that we consider calibration at both the first and second phases, thus allowing for the development of the technique for chain ratio and chain regression type estimators. We also investigate, via a simulation study, the efficiency of the jackknife estimators of variance relative to the usual estimators.

## 2. Estimation of variance using jackknifing

In what follows, we assume that a single stage design is employed at both of the two phases in the sampling process.

Let  $\hat{Y}^c(j)$  be a calibrated estimator of the population mean,  $\bar{Y}$ , obtained by dropping the  $j^{\text{th}}$  unit from the sample  $s_1$  of  $m$  units. We prove in the Appendix that the jackknife estimator of the population mean in two phase-sampling can be written as

$$\hat{Y}^c(j) = \begin{cases} \hat{Y}_2^o(j) + \hat{\beta}_2(j) \{ \hat{X}_1^o(j) - \hat{X}_2^o(j) \} \\ + \hat{\beta}_1(j) \hat{\beta}_2(j) \{ \bar{Z} - \hat{Z}_1^o(j) \} & \text{if } j \in s_2 \\ \hat{Y}_2^o + \hat{\beta}_2 \{ \hat{X}_1^o(j) - \hat{X}_1^o \} \\ + \hat{\beta}_1(j) \hat{\beta}_2 \{ \bar{Z} - \hat{Z}_1^o(j) \} & \text{if } j \in (s_1 - s_2) \end{cases} \quad (2.1)$$

where the quantity  $\hat{Z}_1^o(j) = \hat{Z}_1^o + \{w_{1j}^o / (1 - w_{1j}^o)\} \{ \hat{Z}_1^o - z_j \}$ , the terms  $\hat{X}_1^o(j)$ ,  $\hat{X}_2^o(j)$ , and  $\hat{Y}_2^o(j)$  are defined in an analogous manner,  $\hat{\beta}_1(j) = \hat{\beta}_1 + \{q_{1j} w_{1j}^o z_j (x_j - \hat{\beta}_1 z_j)\} / \{q_{1j} w_{1j}^o z_j^2 - \sum_{i \in s_1} q_{1i} w_{1i}^o z_i^2\}$ , and  $\hat{\beta}_2(j) = \hat{\beta}_2 + \{q_{2j} w_{2j}^o x_j (y_j - \hat{\beta}_2 x_j)\} / \{q_{2j} w_{2j}^o x_j^2 - \sum_{i \in s_1} q_{2i} w_{2i}^o x_i^2\}$ . The modified jackknife estimator of variance of  $\hat{Y}^c$  is then given by

$$\hat{V}_{\text{JACK}}(\hat{Y}^c) = \{(m-1)/m\} \sum_{j \in s_1} \{ \hat{Y}^c(j) - \hat{Y}^c \}^2. \quad (2.2)$$

We show in the appendix that this estimator is consistent.

Note that we can write that

$$\hat{Y}^c(j) - \hat{Y}^c = \begin{cases} \varepsilon_2(j) + \hat{\beta}_2 \varepsilon_1(j) + \hat{\beta}_2(j) d_2(j) \\ + \hat{\beta}_2 \delta_2(j) & \text{if } j \in s_2 \\ \hat{\beta}_2 \varepsilon_1(j) & \text{if } j \in (s_1 - s_2) \end{cases} \quad (2.3)$$

where the terms in (2.3) are given by  $\varepsilon_1(j) = \{ \hat{X}_1^o(j) - \hat{X}_1^o \} - \hat{\beta}_1(j) \{ \hat{Z}_1^o(j) - \bar{Z} \}$ ,  $\varepsilon_2(j) = \{ \hat{Y}_2^o(j) - \hat{Y}_2^o \} - \hat{\beta}_2(j) \{ \hat{X}_2^o(j) - \hat{X}_2^o \} - \hat{\beta}_1(j) \hat{\beta}_2(j) \{ \hat{Z}_1^o(j) - \bar{Z} \}$ ,  $d_2(j) = \{ \hat{X}_1^o(j) - \hat{X}_2^o(j) \}$  and  $\delta_2(j) = \{ \hat{X}_2^o(j) - \hat{X}_1^o \} - \hat{\beta}_1(j) \{ \bar{Z} - \hat{Z}_1^o(j) \} - \hat{\beta}_1 \{ \bar{Z} - \hat{Z}_1^o \}$ . The  $\varepsilon_1(j)$  term is analogous to the error term associated with the regression of the auxiliary variable  $x_i$  on  $z_i$ , for  $i \in s_1$ , while  $\varepsilon_2(j)$  is analogous to the error term associated with the regression of the study variable  $y_i$  on both  $x_i$  and  $z_i$  simultaneously, for  $i \in s_2$ . Provided that  $j \in s_2$ , the  $d_2(j)$  term reflects the difference in the jackknife first and second phase sample means for the variable  $X$ , while  $\delta_2(j)$  denotes an adjustment to  $d_2(j)$  obtained by using information on the auxiliary variable  $Z$ .

Using (2.3) in (2.2), the jackknife estimator of variance of the estimator  $\hat{Y}^c$  is given by

$$\begin{aligned} \hat{V}_{\text{JACK}}(\hat{Y}^c) &= \{(m-1)/m\} \\ &\left[ \sum_{j \in s_2} \varepsilon_2^2(j) + \sum_{j \in s_2} \hat{\beta}_2^2(j) d_2^2(j) \right. \\ &\quad + \hat{\beta}_2^2 \sum_{j \in s_2} \delta_2(j) \{ \delta_2(j) + 2\varepsilon_1(j) \} \\ &\quad + 2\hat{\beta}_2 \sum_{j \in s_2} \varepsilon_1(j) \varepsilon_2(j) \\ &\quad + 2\hat{\beta}_2 \sum_{j \in s_2} \hat{\beta}_2(j) d_2(j) \{ \varepsilon_1(j) + \delta_2(j) \} \\ &\quad \left. + \hat{\beta}_2^2 \sum_{j \in s_1} \varepsilon_1^2(j) \right]. \end{aligned} \quad (2.4)$$

Note that the expression given in (2.4) is exact. It can be used to estimate the variance of several estimators available in the literature.

### 3. Special cases

In the next section, we demonstrate that the estimator proposed by Rao and Sitter (1995), Sitter (1997), Raj (1965), Srivenkataramana and Tracy (1989), Chand (1975), and Ahmed (1997) can be viewed as special cases of the proposed technique.

#### Case 3.1: Rao and Sitter (1995)

If  $\hat{X}_1^c = \hat{X}_1^o$  (no first-phase calibration is made) and  $q_{2i} = 1/x_i$ , then the calibrated estimator of  $\bar{Y}$  becomes

$$\hat{Y}_r^c = \left( \sum_{i \in s_2} w_{2i}^o y_i \right) \left\{ \left( \sum_{i \in s_2} w_{1i}^o x_i \right) / \left( \sum_{i \in s_2} w_{2i}^o x_i \right) \right\}.$$

If the first-phase sample  $s_1$  is selected according to SRSWOR such that the first-phase design weights are given by  $d_{1i} = N/m$ , and the second-phase sample  $s_2$  is selected from  $s_1$  by SRSWOR such that  $d_{2i} = m/n$ , then the calibrated estimator of the population mean becomes

$$\hat{Y}_{\text{RS}}^c = \bar{y}(\bar{x}' / \bar{x}), \quad (3.1)$$

where  $\bar{y} = \sum_{i \in s_2} y_i / n$ ,  $\bar{x} = \sum_{i \in s_2} x_i / n$ , and  $\bar{x}' = \sum_{i \in s_1} x_i / m$ . The jackknife mechanism in (2.1) becomes

$$\hat{Y}_{\text{RS}}^c(j) = \begin{cases} \frac{(n\bar{y} - y_j)(m\bar{x}' - x_j)}{(n\bar{x} - x_j)(m-1)} & \text{if } j \in s_2 \\ \frac{(\bar{y} / \bar{x})(m\bar{x}' - x_j)}{(m-1)} & \text{if } j \in (s_1 - s_2). \end{cases} \quad (3.2)$$

Setting  $\hat{R} = \bar{y} / \bar{x}$ , the difference between (3.2) and (3.1) can be written as



$$\hat{Y}_{RS}^c(j) - \hat{Y}_{RS}^c = \begin{cases} -\hat{R} \frac{(x_j - \bar{x}')}{(m-1)} - \frac{\bar{x}'(j)}{\bar{x}(j)} \frac{(y_j - \hat{R}x_j)}{(n-1)} & \text{if } j \in s_2 \\ -\hat{R} \frac{(x_j - \bar{x}')}{(m-1)} & \text{if } j \in (s_1 - s_2). \end{cases} \quad (3.3)$$

Expression (3.3) is exactly the same as reported by Rao and Sitter (1995). Assuming that  $\bar{x}'(j)/\bar{x}(j) \approx \bar{x}'/\bar{x}$ , then the approximate jackknife estimator of variance is given by

$$\hat{V}_{JACK}(\hat{Y}_{RS}^c) \approx \left( \frac{\bar{x}'}{\bar{x}} \right)^2 \sum_{i \in s_2} \frac{(y_i - \hat{R}x_i)^2}{n(n-1)} + 2 \left( \frac{\bar{x}'}{\bar{x}} \right) \hat{R} \sum_{j \in s_2} \frac{(x_j - \bar{x}')(y_j - \hat{R}x_j)}{n-1} + \hat{R}^2 \sum_{j \in s_2} \frac{(x_j - \bar{x}')^2}{m(m-1)}.$$

Thus, the Rao and Sitter (1995) estimator is a special case of the proposed jackknife technique.

### Case 3.2: Sitter (1997)

In Case 3.1, if we consider  $q_{2i} = 1$ , then the calibrated estimator under SRSWOR becomes

$$\hat{Y}_{lr}^c = \bar{y} + b^*(\bar{x}' - \bar{x}), \quad (3.4)$$

where  $b^* = \sum_{i \in s_2} x_i y_i / \sum_{i \in s_2} x_i^2$  denotes an estimator of the regression coefficient  $\beta$  that is slightly different from the one considered by Sitter (1997). The jackknife mechanism takes the form

$$\hat{Y}_{lr}^c(j) = \begin{cases} \frac{n\bar{y} - y_j}{n-1} + \left\{ b^* + \frac{x_j(y_j - b^*x_j)}{\sum_{i \in s_2} x_i^2 - x_j^2} \right\} \\ \left\{ \frac{m\bar{x}' - x_j}{m-1} - \frac{n\bar{x} - x_j}{n-1} \right\} & \text{if } j \in s_2 \\ \bar{y} + b^* \left\{ \frac{m\bar{x}' - x_j}{m-1} - \bar{x} \right\} & \text{if } j \in (s_1 - s_2). \end{cases} \quad (3.5)$$

If we set  $d_j^* = (y_j - \bar{y}) - b^*(x_j - \bar{x})$ ,  $a_j^* = x_j\{\bar{x}(j) - \bar{x}'(j)\}/K$ , and  $k_j^* = x_j^2/K$ , where  $K = (n-1)s_x^2 + n\bar{x}^2$ , then the difference between (3.5) and (3.4) can be written as

$$\hat{Y}_{lr}^c(j) - \hat{Y}_{lr}^c = \begin{cases} -b^* \frac{(x_j - \bar{x}')}{(m-1)} - \frac{\frac{d_j^*}{(n-1)}}{\left[ 1 + \frac{a_j^*}{(1-k_j^*)} \right]} & \text{if } j \in s_2 \\ -b^* \frac{(x_j - \bar{x}')}{(m-1)} & \text{if } j \in (s_1 - s_2) \end{cases}$$

which is similar to the expression reported by Sitter (1997).

### Case 3.3: Raj (1965)

In order to consider this case, we assume that the initial sample  $s_1$  of size  $m$  is selected with replacement according to probabilities  $p_i$  proportional to  $z_i$ ,  $i = 1, 2, \dots, N$ . Information on the auxiliary variable  $X$  is collected on this first-phase sample,  $s_1$ . The second-phase sample, specified to be of size  $n$ , is a subsample of  $s_1$  selected without replacement using equal probabilities. It is for  $s_2$  that information on  $Y$  is collected. Under this sampling scheme,  $d_{1i} = 1/\pi_{1i} = 1/(mp_i)$  and  $d_{2i} = m/n$ . Thus,  $w_{1i}^o = (1/p_i)/\sum_{i \in s_1} (1/p_i)$  and  $w_{2i}^o = (1/p_i)/\sum_{i \in s_2} (1/p_i)$ . Note also that for this scheme,  $\hat{X}_1^c = \hat{X}_1^o$ ; thus no first-phase calibration is made. If  $q_{2i} = 1/x_i$ , then the calibrated estimator  $\hat{Y}^c$  becomes

$$\hat{Y}_{Raj}^c = \hat{Y}_2^o (\hat{X}_1^o / \hat{X}_2^o), \quad (3.6)$$

where  $\hat{Y}_2^o = \sum_{i \in s_2} (y_i/p_i)/\sum_{i \in s_2} (1/p_i)$ ,  $\hat{X}_2^o = \sum_{i \in s_2} (x_i/p_i)/\sum_{i \in s_2} (1/p_i)$ , and  $\hat{X}_1^o = \sum_{i \in s_1} (x_i/p_i)/\sum_{i \in s_1} (1/p_i)$ . Thus, alternatively  $\hat{Y}_{Raj}^c = \{\sum_{i \in s_2} (y_i/p_i)\sum_{i \in s_1} (x_i/p_i)\} / \{\sum_{i \in s_2} (x_i/p_i)\sum_{i \in s_1} (1/p_i)\}$ .

Under the sampling scheme described above, the jackknife estimator of population mean is

$$\hat{Y}_{Raj}^c(j) = \begin{cases} \hat{Y}_2^o(j) \frac{\hat{X}_1^o(j)}{\hat{X}_2^o(j)} & \text{if } j \in s_2 \\ \hat{Y}_2^o \frac{\hat{X}_1^o(j)}{\hat{X}_2^o} & \text{if } j \in (s_1 - s_2) \end{cases} \quad (3.7)$$

where

$$\hat{Y}_2^o(j) = \frac{\sum_{i \in s_2} (y_i/p_i)}{\sum_{i \in s_2} (1/p_i)} + \frac{(1/p_j)/\sum_{i \in s_2} (1/p_i)}{1 - \frac{(1/p_j)}{\sum_{i \in s_2} (1/p_i)}} \left\{ \frac{\sum_{i \in s_2} (y_i/p_i)}{\sum_{i \in s_2} (1/p_i)} - y_j \right\},$$

and  $\hat{X}_2^o(j)$  and  $\hat{X}_1^o(j)$  are defined analogously. If  $\hat{R} = \hat{Y}_2^o / \hat{X}_2^o$  and  $w_{2j}^o = (1/p_j) / \sum_{i \in s_2} (1/p_i)$ , the difference between (3.7) and (3.6) can easily be written as

$$\hat{Y}_{\text{Raj}}^c(j) - \hat{Y}_{\text{Raj}}^c = \begin{cases} -w_{2j}^o \frac{\hat{X}_1^o(j)}{\hat{X}_2^o(j)} (y_j - \hat{R}x_j) \\ \quad + \hat{R} \{ \hat{X}_1^o(j) - \hat{X}_1^o \} & \text{if } j \in s_2 \\ \hat{R} \{ \hat{X}_1^o(j) - \hat{X}_1^o \} & \text{if } j \in (s_1 - s_2). \end{cases}$$

Thus, the jackknife estimator of variance of the estimator  $\hat{Y}_{\text{Raj}}^c$  is given by

$$\begin{aligned} \hat{V}_{\text{JACK}}(\hat{Y}_{\text{Raj}}^c) = & \frac{m-1}{m} \left[ \sum_{j \in s_2} (w_{2j}^o)^2 \frac{\hat{X}_1^o(j)^2}{\hat{X}_2^o(j)^2} (y_j - \hat{R}x_j)^2 \right. \\ & + \hat{R}^2 \sum_{j \in s_1} \{ \hat{X}_1^o(j) - \hat{X}_1^o \}^2 \\ & \left. - 2\hat{R} \sum_{j \in s_2} w_{2j}^o \frac{\hat{X}_1^o(j)}{\hat{X}_2^o(j)} (y_j - \hat{R}x_j) \{ \hat{X}_1^o(j) - \hat{X}_1^o \} \right]. \end{aligned}$$

Following Rao and Sitter (1995), if we assume  $\hat{X}_1^o(j) / \hat{X}_2^o(j) \approx \hat{X}_1^o / \hat{X}_2^o$ , then the jackknife estimator of variance of  $\hat{Y}_{\text{Raj}}^c$  takes the form

$$\begin{aligned} \hat{V}_{\text{JACK}}(\hat{Y}_{\text{Raj}}^c) \approx & \frac{m-1}{m} \left[ \{ \hat{X}_1^o / \hat{X}_2^o \}^2 \sum_{j \in s_2} (w_{2j}^o)^2 (y_j - \hat{R}x_j)^2 \right. \\ & + \hat{R}^2 \sum_{j \in s_1} \{ \hat{X}_1^o(j) - \hat{X}_1^o \}^2 \\ & \left. - 2\hat{R} \{ \hat{X}_1^o / \hat{X}_2^o \} \sum_{j \in s_2} w_{2j}^o (y_j - \hat{R}x_j) \{ \hat{X}_1^o(j) - \hat{X}_1^o \} \right]. \end{aligned}$$

### Case 3.4: Srivenkataramana and Tracy (1989)

In order to consider this case, as in Raj (1965), we assume that the initial sample  $s_1$  of size  $m$  is selected with replacement according to probabilities proportional to  $z_i$ . However, the subsample,  $s_2$ , of  $n$  units is now selected with replacement using probabilities proportional to  $x_i / z_i$ . As a result,  $w_{1i}^o = (1/z_i) / \sum_{i \in s_1} (1/z_i)$  and  $w_{2i}^o = (1/x_i) / \sum_{i \in s_2} (1/x_i)$ . Similar to Raj (1965), no first-phase calibration is made; thus  $\hat{X}_1^c = \hat{X}_1^o$ . Hence, if  $q_{2i} = 1/x_i$ , then the calibrated estimator  $\hat{Y}^c$  is

$$\hat{Y}_{\text{ST}}^c = \hat{Y}_2^o (\hat{X}_1^o / \hat{X}_2^o), \quad (3.8)$$

where  $\hat{Y}_2^o = \sum_{i \in s_2} (y_i / x_i) / \sum_{i \in s_2} (1/x_i)$ ,  $\hat{X}_2^o = n / \sum_{i \in s_2} (1/x_i)$ , and  $\hat{X}_1^o = \sum_{i \in s_1} (x_i / z_i) / \sum_{i \in s_1} (1/z_i)$ . Thus, alternatively  $\hat{Y}_{\text{ST}}^c = \{ \sum_{i \in s_2} (y_i / x_i) \sum_{i \in s_1} (x_i / z_i) \} / \{ n \sum_{i \in s_1} (1/z_i) \}$ .

Under the sampling scheme described above, the jackknife estimator of population mean is

$$\hat{Y}_{\text{ST}}^c(j) = \begin{cases} \hat{Y}_2^o(j) \{ \hat{X}_1^o(j) / \hat{X}_2^o(j) \} & \text{if } j \in s_2 \\ \hat{Y}_2^o \{ \hat{X}_1^o(j) / \hat{X}_2^o \} & \text{if } j \in (s_1 - s_2) \end{cases} \quad (3.9)$$

where

$$\hat{Y}_2^o(j) = \frac{\sum_{i \in s_2} (y_i / x_i)}{\sum_{i \in s_2} (1/x_i)} + \frac{1}{x_j \sum_{i \in s_2} (1/x_i) - 1} \left\{ \frac{\sum_{i \in s_2} (y_i / x_i)}{\sum_{i \in s_2} (1/x_i)} - y_j \right\}.$$

The terms  $\hat{X}_2^o(j)$  and  $\hat{X}_1^o(j)$  are defined similarly; that is

$$\hat{X}_2^o(j) = \frac{n}{\sum_{i \in s_2} (1/x_i)} + \frac{1}{x_j \sum_{i \in s_2} (1/x_i) - 1} \left\{ \frac{n}{\sum_{i \in s_2} (1/x_i)} - x_j \right\},$$

while  $\hat{X}_1^o(j)$  can be written as

$$\hat{X}_1^o(j) = \frac{\sum_{i \in s_1} (x_i / z_i)}{\sum_{i \in s_1} (1/z_i)} + \frac{1}{x_j \sum_{i \in s_1} (1/z_i) - 1} \left\{ \frac{\sum_{i \in s_1} (x_i / z_i)}{\sum_{i \in s_1} (1/z_i)} - x_j \right\}.$$

If  $\hat{R} = \sum_{i \in s_2} (y_i / x_i) / n$  and  $w_{2j}^o = (1/x_j) / \sum_{i \in s_2} (1/x_i)$ , the difference between (3.9) and (3.8) is given by

$$\begin{aligned} \hat{Y}_{\text{ST}}^c(j) - \hat{Y}_{\text{ST}}^c = & \begin{cases} -w_{2j}^o \frac{\hat{X}_1^o(j)}{\hat{X}_2^o(j)} (y_j - \hat{R}x_j) + \hat{R} \{ \hat{X}_1^o(j) - \hat{X}_1^o \} & \text{if } j \in s_2 \\ \hat{R} \{ \hat{X}_1^o(j) - \hat{X}_1^o \} & \text{if } j \in (s_1 - s_2). \end{cases} \end{aligned}$$

Following Rao and Sitter (1995), if we assume  $\hat{X}_1^o(j) / \hat{X}_2^o(j) \approx \hat{X}_1^o / \hat{X}_2^o$ , then the jackknife estimator of variance of  $\hat{Y}_{\text{ST}}^c$  takes the form

$$\begin{aligned} \hat{V}_{\text{JACK}}(\hat{Y}_{\text{ST}}^c) \approx & \frac{m-1}{m} \left[ \{ \hat{X}_1^o / \hat{X}_2^o \}^2 \sum_{j \in s_2} (w_{2j}^o)^2 (y_j - \hat{R}x_j)^2 \right. \\ & + \hat{R}^2 \sum_{j \in s_1} \{ \hat{X}_1^o(j) - \hat{X}_1^o \}^2 \\ & \left. - 2\hat{R} \{ \hat{X}_1^o / \hat{X}_2^o \} \sum_{j \in s_2} w_{2j}^o (y_j - \hat{R}x_j) \{ \hat{X}_1^o(j) - \hat{X}_1^o \} \right]. \end{aligned}$$

### Case 3.5: Chand (1975)

In order to consider this case, the first-phase sample  $s_1$  of size  $m$  is selected using SRSWOR, and both auxiliary



variables  $Z$  and  $X$  are observed on the chosen units. The subsample,  $s_2$ , of  $n$  units is also selected using SRSWOR. Obviously,  $d_{1i} = N/m$  and  $d_{2i} = m/n$ , so that  $w_{1i}^o = 1/m$  and  $w_{2i}^o = 1/n$ . If  $q_{1i} = 1/z_i$  and  $q_{2i} = 1/x_i$ , then the calibrated estimator  $\hat{Y}_{Ch}^c$  becomes

$$\hat{Y}_{Ch}^c = \bar{y}(\bar{x}'/\bar{x})(\bar{Z}/\bar{z}'), \quad (3.10)$$

where

$$\bar{y} = \sum_{i \in s_2} y_i / n, \quad \bar{x} = \sum_{i \in s_2} x_i / n, \quad \bar{x}' = \sum_{i \in s_2} x_i / m,$$

and  $\bar{z}' = \sum_{i \in s_1} z_i / m$ . The jackknife estimator of  $\bar{Y}$  is

$$\hat{Y}_{Ch}^c(j) = \begin{cases} \bar{y}(j) \frac{\bar{x}'(j)}{\bar{x}(j)} \frac{\bar{Z}}{\bar{z}'(j)} & \text{if } j \in s_2 \\ \bar{y}(j) \frac{\bar{x}'(j)}{\bar{x}} \frac{\bar{Z}}{\bar{z}'(j)} & \text{if } j \in (s_1 - s_2) \end{cases} \quad (3.11)$$

where  $\bar{y}(j) = (n\bar{y} - y_j)/(n-1)$ ,  $\bar{x}(j) = (n\bar{x} - x_j)/(n-1)$ ,  $\bar{x}'(j) = (m\bar{x}' - x_j)/(m-1)$ , and finally  $\bar{z}'(j) = (m\bar{z}' - z_j)/(m-1)$ . If we let  $\hat{R}_1 = \bar{x}'/\bar{z}'$  (an estimator of  $R_1 = \bar{X}/\bar{Z}$ ) and  $\hat{R}_2 = \bar{y}/\bar{x}$  (an estimator of  $R_2 = \bar{Y}/\bar{X}$ ), and similarly, let  $\hat{R}_1(j) = \bar{x}'(j)/\bar{z}'(j)$  and  $\hat{R}_2(j) = \bar{y}(j)/\bar{x}(j)$ , the difference between (3.11) and (3.10) can be written as

$$\hat{Y}_{Ch}^c(j) - \hat{Y}_{Ch}^c = \begin{cases} \varepsilon_2(j) + \hat{R}_2 \varepsilon_1(j) + \hat{R}_2(j) d_2(j) + \hat{R}_2 \delta_2(j) & \text{if } j \in s_2 \\ \hat{R}_2 \varepsilon_1(j) & \text{if } j \in (s_1 - s_2) \end{cases} \quad (3.12)$$

where we can write in (3.12) that  $\varepsilon_2(j) = \{\bar{y}(j) - \bar{y}\} - \hat{R}_2(j) \{\bar{x}(j) - \bar{x}\} - \hat{R}_1(j) \hat{R}_2(j) \{\bar{z}'(j) - \bar{Z}\}$ ,  $d_2(j) = \{\bar{x}'(j) - \bar{x}'\}$ ,  $\delta_2(j) = \{\bar{x}(j) - \bar{x}'(j)\} - \hat{R}_1(j) \{\bar{Z} - \bar{z}'(j)\} - \hat{R}_1 \{\bar{Z} - \bar{z}'\}$ , and finally that the term  $\varepsilon_1(j) = \{\bar{x}'(j) - \bar{x}'\} - \hat{R}_1(j) \{\bar{z}'(j) - \bar{Z}\}$ . Thus the jackknife estimator of variance of the estimator  $\hat{Y}_{Ch}^c$  is given by

$$\begin{aligned} \hat{V}_{JACK}(\hat{Y}_{Ch}^c) = & \{(m-1)/m\} \left[ \sum_{j \in s_2} \varepsilon_2^2(j) + \sum_{j \in s_2} \hat{R}_2^2(j) d_2^2(j) \right. \\ & + \hat{R}_2^2 \sum_{j \in s_2} \delta_2(j) \{\delta_2(j) + 2\varepsilon_1(j)\} \\ & + 2\hat{R}_2 \sum_{j \in s_2} \varepsilon_1(j) \varepsilon_2(j) \\ & + 2\hat{R}_2 \sum_{j \in s_2} \hat{R}_2(j) d_2(j) \{\varepsilon_1(j) + \delta_2(j)\} \\ & \left. + \hat{R}_2^2 \sum_{j \in s_1} \varepsilon_1^2(j) \right]. \end{aligned}$$

### Case 3.6: Ahmed (1997)

Consider the same sample design as in Case 3.5. Rather than  $q_{1i} = 1/z_i$  and  $q_{2i} = 1/x_i$  as in Chand (1975), we set  $q_{1i} = q_{2i} = 1$ , and  $q_{2i} = 1/x_i$ , then the calibrated estimator reduces to

$$\hat{Y}_{Chlr}^c = \bar{y} + b_2^*(\bar{x}' - \bar{x}) + b_1^* b_2^*(\bar{Z} - \bar{z}'), \quad (3.13)$$

where  $b_2^* = \sum_{i \in s_2} x_i y_i / \sum_{i \in s_2} x_i^2$  and  $b_1^* = \sum_{i \in s_1} x_i z_i / \sum_{i \in s_1} z_i^2$ . Note that (3.13) is a chain regression type estimator similar to Ahmed (1997). Letting  $b_2^*(j) = b_2^* + \{x_j(y_j - b_2^* x_j) / (x_j^2 - \sum_{i \in s_2} x_i^2)\}$  and  $b_1^*(j) = b_1^* + \{z_j(x_j - b_1^* z_j) / (z_j^2 - \sum_{i \in s_1} z_i^2)\}$ , after jackknifing the estimator  $\hat{Y}_{Chlr}^c$  becomes

$$\hat{Y}_{Chlr}^c(j) = \begin{cases} \bar{y}(j) + b_2^*(j) \{\bar{x}'(j) - \bar{x}(j)\} \\ + b_1^*(j) b_2^*(j) \{\bar{Z} - \bar{z}'(j)\} & \text{if } j \in s_2 \\ \bar{y} + b_2^* \{\bar{x}'(j) - \bar{x}\} \\ + b_1^*(j) b_2^* \{\bar{Z} - \bar{z}'(j)\} & \text{if } j \in (s_1 - s_2). \end{cases} \quad (3.14)$$

The difference between (3.14) and (3.13) can be written as

$$\hat{Y}_{Chlr}^c(j) - \hat{Y}_{Chlr}^c = \begin{cases} \varepsilon_2(j) + b_2^* \varepsilon_1(j) + b_2^*(j) d_2(j) + b_2^* \delta_2(j) & \text{if } j \in s_2 \\ b_2^* \varepsilon_1(j) & \text{if } j \in (s_1 - s_2) \end{cases} \quad (3.15)$$

where we can write in (3.15) that  $\varepsilon_2(j) = \{\bar{y}(j) - \bar{y}\} - b_2^*(j) \{\bar{x}(j) - \bar{x}\} - b_1^*(j) b_2^*(j) \{\bar{z}'(j) - \bar{Z}\}$ ,  $d_2(j) = \{\bar{x}'(j) - \bar{x}'\}$ ,  $\delta_2(j) = \{\bar{x}(j) - \bar{x}'(j)\} - b_1^*(j) \{\bar{Z} - \bar{z}'(j)\} - b_1^* \{\bar{Z} - \bar{z}'\}$ , and finally that the term  $\varepsilon_1(j) = \{\bar{x}'(j) - \bar{x}'\} - b_1^*(j) \{\bar{z}'(j) - \bar{Z}\}$ . Thus the jackknife estimator of variance of the estimator  $\hat{Y}_{Chlr}^c$  is given by

$$\begin{aligned} \hat{V}_{JACK}(\hat{Y}_{Chlr}^c) = & \{(m-1)/m\} \left[ \sum_{j \in s_2} \varepsilon_2^2(j) + \sum_{j \in s_2} \{b_2^*(j)\}^2 d_2^2(j) \right. \\ & + \{b_2^*(j)\}^2 \sum_{j \in s_2} \delta_2(j) \{\delta_2(j) + 2\varepsilon_1(j)\} \\ & + 2b_2^* \sum_{j \in s_2} \varepsilon_1(j) \varepsilon_2(j) \\ & + 2b_2^* \sum_{j \in s_2} b_2^*(j) d_2(j) \{\varepsilon_1(j) + \delta_2(j)\} \\ & \left. + \{b_2^*\}^2 \sum_{j \in s_1} \varepsilon_1^2(j) \right]. \end{aligned}$$

## 4. Simulation study

In this section, we present the results of simulation studies designed to investigate the performance of the proposed jackknife procedure for estimating the variance of four of the two-phase estimators of population mean

presented in Section 3. Specifically, we consider the Rao and Sitter (1995) ratio-type estimator, the Sitter (1997) regression-type estimator, the Chand (1975) chain ratio-type estimator, and the Ahmed (1997) chain regression-type estimator. Initially, we describe and report the results of simulations that were conducted for the Sitter and Rao (1995) and Sitter (1997) estimators. This is followed by a discussion and summary of similar simulations on the Chand (1975) and Ahmed (1997) estimators. Unlike the case for the ratio and regression estimators, since complete information on a second auxiliary variable  $Z$  is required for the entire population in order to apply the two chain estimators, the simulations that were conducted for these two estimators are somewhat more complicated than those performed for the ratio and regression estimators.

#### 4.1 Simulation study: Rao and Sitter (1995) and Sitter (1997)

For purposes of the first set of simulations, we assume that a first-phase sample of  $m$  units is selected from a population of  $N$  units, and only the auxiliary variable  $X$  is measured. From the first-phase sample of  $m$  units, we then select a second-phase sample of  $n$  units by SRSWOR in which both the study variable,  $Y$ , and the auxiliary variable,  $X$ , are measured.

We began by creating a population of  $N$  units consisting of  $(X_i, Y_i)$  pairs using the model

$$Y_i = \beta X_i + \sqrt{X_i^g} \varepsilon_i,$$

with  $\beta = 10$ . Initially, we set  $g = 0$  and  $N = 500$ . For each  $i, i = 1, \dots, N$ , we generated  $X_i$  from a gamma distribution with a shape parameter of 3.1 and a scale parameter of one, and  $\varepsilon_i$  from a standard normal. From the resulting population of  $(X_i, Y_i)$  pairs, we selected 1,000 first-phase sample of  $m = 100$  units, and from each of these samples, we selected 10,000 second-phase samples of  $n = 20$  units.

Under the sampling scheme used here, Rao and Sitter (1995) proposed the ratio estimator

$$\hat{Y}_{RS}^c = \bar{y}(\bar{x}' / \bar{x}), \quad (4.1)$$

which has approximate variance

$$V(\hat{Y}_{RS}^c) = (n^{-1} - m^{-1})S_d^2 + (m^{-1} - N^{-1})S_y^2,$$

where

$$S_d^2 = (N - 1)^{-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - R(X_i - \bar{X})]^2$$

and

$$S_y^2 = (N - 1)^{-1} \sum_{i=1}^N (Y_i - \bar{Y})^2,$$

with  $\bar{Y} = \sum_{i=1}^N Y_i / N$ ,  $\bar{X} = \sum_{i=1}^N X_i / N$ , and  $R = \bar{Y} / \bar{X}$ . For the  $t^{\text{th}}$  second phase sample ( $t = 1, \dots, 10,000$ ) drawn from the  $k^{\text{th}}$  first phase sample ( $k = 1, \dots, 1,000$ ), we computed the usual estimator of variance

$$\hat{V}[(\hat{Y}_{RS}^c(t|k))] = \left(\frac{1}{n} - \frac{1}{m}\right) s_{d(t|k)}^2 + \left(\frac{1}{m} - \frac{1}{N}\right) s_{y(t|k)}^2, \quad (4.2)$$

where the sample variances are

$$s_{d(t|k)}^2 = (n - 1)^{-1} \sum_{i=1}^n [(y_{i(t|k)} - \bar{y}_{(t|k)}) - r_{(t|k)}(x_{i(t|k)} - \bar{x}_{(t|k)})]^2$$

and

$$s_{y(t|k)}^2 = (n - 1)^{-1} \sum_{i=1}^n (y_{i(t|k)} - \bar{y}_{(t|k)})^2$$

with  $\bar{y}_{(t|k)} = \sum_{i=1}^n y_{i(t|k)} / n$  and  $\bar{x}_{(t|k)} = \sum_{i=1}^n x_{i(t|k)} / n$ . In addition,  $r_{(t|k)} = \bar{y}_{(t|k)} / \bar{x}_{(t|k)}$ . We also computed the jackknife estimator of variance

$$\begin{aligned} \hat{V}_{JACK}[(\hat{Y}_{RS}^c(t|k))] = \\ \frac{m-1}{m} \sum_{j=1}^m \left[ \bar{y}_{(t|k)}(j) \frac{\bar{x}'_{(t|k)}(j)}{\bar{x}_{(t|k)}(j)} - \bar{y}_{(t|k)} \frac{\bar{x}'_{(t|k)}}{\bar{x}_{(t|k)}} \right]^2, \end{aligned} \quad (4.3)$$

and the ratio of estimated variances

$$RV(t|k) = \hat{V}[(\hat{Y}_{RS}^c(t|k))] / \hat{V}_{JACK}[(\hat{Y}_{RS}^c(t|k))].$$

We then computed the average of the  $RV(t|k)$  over all  $k$  and  $t$ , which is given by

$$RV = \frac{1}{10,000,000} \sum_{k=1}^{1,000} \sum_{t=1}^{10,000} RV(t|k).$$

We also determined empirical estimates of the biases in (4.2) and (4.3) by computing

$$EBU = \frac{1}{10,000,000} \sum_{k=1}^{1,000} \sum_{t=1}^{10,000} \{\hat{V}[(\hat{Y}_{RS}^c(t|k))] - V(\hat{Y}_{RS}^c)\},$$

and

$$EBJ = \frac{1}{10,000,000} \sum_{k=1}^{1,000} \sum_{t=1}^{10,000} \{\hat{V}_{JACK}[(\hat{Y}_{RS}^c(t|k))] - V(\hat{Y}_{RS}^c)\}.$$

Note that the estimator given in (4.2) is unbiased. Finally, we calculated the relative efficiency of the usual estimator of variance to the jackknife estimator according to

$$RE = \left( \frac{\sum_{k=1}^{1,000} \sum_{t=1}^{10,000} \{\hat{V}[(\hat{Y}_{RS}^c(t|k))] - V(\hat{Y}_{RS}^c)\}^2}{\sum_{k=1}^{1,000} \sum_{t=1}^{10,000} \{\hat{V}_{JACK}[(\hat{Y}_{RS}^c(t|k))] - V(\hat{Y}_{RS}^c)\}^2} \right).$$



Using the same generated population of  $N = 500$ , we repeated the simulation; however we used  $m = 400$  and  $n = 80$  instead. We then created four additional populations of size  $N = 500$  using  $g = 0.5, 1.0, 1.5$ , and  $2.0$ . For each of these four populations, we repeated the two simulations described above where in the first simulation,  $m = 100$  with  $n = 20$ , and in the second simulation,  $m = 400$  and  $n = 80$ . Finally, to study the effect of population size, we then repeated all the simulations based on the different values of  $g$ ,  $m$ , and  $n$  when  $N = 500$  for three additional values of  $N$ , namely  $5,000$ ,  $50,000$ , and  $500,000$ . The results obtained for RV, EBU, EBJ, and RE for each of these simulations are presented in Table 1.

The results for RE in Table 1 suggest that as the population size  $N$  tends to infinity (as considered by Rao and Sitter 1995), the jackknife estimator of variance remains more efficient than the usual unbiased estimator of variance. It is also the case for very large  $N$  that the values for RV

tend to one. However, considering the cases where  $N = 500$ , if the population size is relatively small, not only are the values for RV noticeably smaller than one, but the jackknife estimator of variance seems to be significantly biased. In addition, the jackknife estimator appears to be much less efficient than the usual unbiased estimator of variance, especially when  $m$  and  $n$  are large. Of note here is the fact that Rao and Sitter (1995) and Sitter (1997) state that it is not clear how to fix the finite population correction factors in the jackknife estimator of variance in two-phase sampling. This would seem to be an area where further research could be fruitful, since it would appear that when the population size is small, it might be worthwhile to adjust the finite population correction factors instead of directly applying the jackknife technique according to the approach proposed here. Note that Kim *et al.* (2006) have incorporated a finite population correction factor in a special case.

**Table 1**

Comparison of the jackknife and usual estimators of variance of the ratio estimator of the population mean when  $\beta = 10$  and the auxiliary variable,  $X$ , follows a gamma distribution with a shape parameter of 3.1 and a scale parameter of one

$N$	$m$	$n$	$g$	RV	EBU	EBJ	RE
500	100	20	0.0	0.801	0.006	0.542	1.521
			0.5	0.800	0.010	0.579	1.310
			1.0	0.805	-0.071	0.561	1.267
			1.5	0.816	-0.358	0.575	1.149
			2.0	0.840	-0.720	1.777	0.935
5,000	100	20	0.0	0.979	-0.028	0.042	4.015
			0.5	0.976	0.007	0.096	3.709
			1.0	0.965	0.023	0.172	3.210
			1.5	0.936	-0.073	0.337	1.308
			2.0	0.916	-1.103	0.493	0.967
50,000	100	20	0.0	1.001	-0.002	0.003	6.241
			0.5	0.998	0.107	0.126	4.936
			1.0	0.981	0.101	0.196	2.965
			1.5	0.937	-0.211	0.167	1.558
			2.0	0.924	-0.355	0.940	1.005
500,000	100	20	0.0	1.001	-0.057	-0.054	4.730
			0.5	0.999	0.014	0.024	4.669
			1.0	0.993	0.185	0.229	3.223
			1.5	0.940	-0.235	0.122	1.420
			2.0	0.907	-1.054	0.530	1.009
500	400	80	0.0	0.214	0.000	0.520	0.002
			0.5	0.237	-0.001	0.523	0.002
			1.0	0.320	0.000	0.544	0.006
			1.5	0.530	-0.001	0.616	0.066
			2.0	0.733	-0.012	1.091	0.452
5,000	400	80	0.0	0.919	-0.003	0.061	2.687
			0.5	0.920	-0.001	0.064	2.505
			1.0	0.922	0.003	0.077	2.058
			1.5	0.930	-0.028	0.077	1.372
			2.0	0.940	-0.089	0.184	1.088
50,000	400	80	0.0	0.991	-0.008	-0.001	4.550
			0.5	0.991	0.004	0.012	5.276
			1.0	0.991	0.000	0.009	4.163
			1.5	0.980	-0.024	-0.001	1.777
			2.0	0.967	-0.171	-0.040	1.099
500,000	400	80	0.0	1.000	0.009	0.009	5.501
			0.5	0.999	0.001	0.001	5.180
			1.0	0.993	-0.001	0.006	3.852
			1.5	0.992	-0.022	-0.018	1.809
			2.0	0.971	-0.179	-0.079	1.136

We also considered the Sitter (1997) regression estimator, and repeated the entire simulation study that was performed using the ratio estimator in (4.1). Specifically, rather than (4.1), we made use of the estimator

$$\hat{Y}_S^c = \bar{y} + b^*(\bar{x}' - \bar{x}), \quad (4.4)$$

which has approximate variance

$$V(\hat{Y}_S^c) = (n^{-1} - m^{-1})S_d^2 + (m^{-1} - N^{-1})S_y^2, \quad (4.5)$$

where

$$S_d^2 = (N-1)^{-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - \beta_{\text{POP}}(X_i - \bar{X})]^2$$

with

$$\beta_{\text{POP}} = \sum_{i=1}^N X_i Y_i / \sum_{i=1}^N X_i^2.$$

For each different combination of  $N$ ,  $g$ ,  $m$ , and  $n$  used in the simulation study, we computed

$$\hat{V}[(\hat{Y}_S^c(t|k))] = (n^{-1} - m^{-1})s_{d(t|k)}^2 + (m^{-1} - N^{-1})s_{y(t|k)}^2, \quad (4.6)$$

for the  $t^{\text{th}}$  second phase sample drawn from the  $k^{\text{th}}$  first phase sample, where the sample variance

$$s_{d(t|k)}^2 = (n-1)^{-1} \sum_{i=1}^n [(y_{i(t|k)} - \bar{y}_{(t|k)}) - b_{(t|k)}^*(x_{i(t|k)} - \bar{x}_{(t|k)})]^2.$$

We also computed the jackknife estimator of variance

$$\begin{aligned} \hat{V}_{\text{JACK}}[(\hat{Y}_S^c(t|k))] = \\ \frac{m-1}{m} \sum_{j=1}^m [\bar{y}_{(t|k)}(j) + b_{(t|k)}^*(j) \{\bar{x}'_{(t|k)}(j) - \bar{x}_{(t|k)}(j)\} \\ - \{\bar{y} + b^*(\bar{x}' - \bar{x})\}]^2. \end{aligned} \quad (4.7)$$

For each different combination of  $N$ ,  $g$ ,  $m$ , and  $n$ , equations (4.5) through (4.7) were used to compute values for RV, EBU, EBJ, and RE analogous to those given in Table 1 for the estimator in (4.1). The results obtained were extremely similar to those for the ratio estimator.

## 4.2 Simulation study: Chand (1975) and Ahmed (1997)

For purposes of the second set of simulations, we now assume that when the first-phase sample of  $m$  units is selected from the population of size  $N$ , information on two auxiliary variables  $X$  and  $Z$  is collected. When the second-phase sample of size  $n$  is selected from the first-phase sample, the study variable  $Y$  is measured, along with the two auxiliary variables  $X$  and  $Z$ . Note also that the

auxiliary variable  $Z$  is assumed to be known for the entire population.

We began by creating a population of  $N = 500$  units of  $(X_i, Z_i, Y_i)$  observations using

$$Y_i = \beta_1 X_i + \beta_2 Z_i + \varepsilon_i,$$

with  $\beta_1 = 3.5$  and  $\beta_2 = 2.5$ . For each  $i$ ,  $i = 1, \dots, N$ , we generated  $X_i$  from a gamma distribution with a shape parameter of 2.2 and a scale parameter of one,  $Z_i$  from a gamma distribution with a shape parameter of 0.1 and a scale parameter of one, and  $\varepsilon_i$  from a standard normal. From the resulting population of  $(X_i, Z_i, Y_i)$  observations, we selected 1,000 first-phase sample of  $m = 100$  units, and from each of these samples, we selected 10,000 second-phase samples of  $n = 20$  units.

Following Chand (1975), a chain ratio estimator under two-phase sampling is given by

$$\hat{Y}_{\text{Ch}}^c = \bar{y}(\bar{x}' / \bar{x})(\bar{Z} / \bar{z}'),$$

which has approximate variance

$$V(\hat{Y}_{\text{Ch}}^c) = (n^{-1} - m^{-1})S_{d_2}^2 + (m^{-1} - N^{-1})S_{d_1}^2, \quad (4.8)$$

where

$$S_{d_2}^2 = (N-1)^{-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - R_2(X_i - \bar{X})]^2$$

and

$$S_{d_1}^2 = (N-1)^{-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - R_1(Z_i - \bar{Z})]^2$$

with

$$\bar{Y} = \sum_{i=1}^N Y_i / N, \quad \bar{X} = \sum_{i=1}^N X_i / N, \quad \bar{Z} = \sum_{i=1}^N Z_i / N,$$

$R_1 = \bar{Y} / \bar{Z}$ , and  $R_2 = \bar{Y} / \bar{X}$ . In the simulation study, we computed

$$\hat{V}[(\hat{Y}_{\text{Ch}}^c(t|k))] = (n^{-1} - m^{-1})s_{d_2(t|k)}^2 + (m^{-1} - N^{-1})s_{d_1(t|k)}^2, \quad (4.9)$$

for the  $t^{\text{th}}$  second phase sample drawn from the  $k^{\text{th}}$  first phase sample, where the sample variances

$$s_{d_2(t|k)}^2 = (n-1)^{-1} \sum_{i=1}^n [(y_{i(t|k)} - \bar{y}_{(t|k)}) - r_{2(t|k)}(x_{i(t|k)} - \bar{x}_{(t|k)})]^2$$

with

$$r_{2(t|k)} = \bar{y}_{(t|k)} / \bar{x}_{(t|k)}$$

and

$$s^2_{d_1(t|k)} = (n-1)^{-1} \sum_{i=1}^n [(y_{i(t|k)} - \bar{y}_{(t|k)}) - r_{1(t|k)}(z_{i(t|k)} - \bar{z}_{(t|k)})]^2$$

with  $r_{1(t|k)} = \bar{y}_{(t|k)} / \bar{z}_{(t|k)}$ . We also computed the jackknife estimator of variance

$$\hat{V}_{\text{JACK}}[(\hat{Y}_{\text{Ch}}^c(t|k))] = \frac{m-1}{m} \sum_{j=1}^m \left[ \bar{y}_{(t|k)}(j) \frac{\bar{x}'_{(t|k)}(j)}{\bar{x}_{(t|k)}(j)} \frac{\bar{Z}}{\bar{z}'_{(t|k)}(j)} - \bar{y}_{(t|k)} \frac{\bar{x}'_{(t|k)}}{\bar{x}_{(t|k)}} \frac{\bar{Z}}{\bar{z}'_{(t|k)}} \right]^2. \tag{4.10}$$

Using the same generated population of  $N = 500$ , we repeated the simulation; however we used  $m = 400$  and  $n = 80$  instead. We then created three additional populations of size  $N = 500$  using  $\beta_1 = 0.5$  with  $\beta_2 = 0.5$ ,  $\beta_1 = 3.5$  with  $\beta_2 = 0.5$ , and  $\beta_1 = 0.5$  with  $\beta_2 = 2.5$ . For each of these three populations, we repeated the two simulations described above where in the first simulation,

$m = 100$  with  $n = 20$ , and in the second simulation,  $m = 400$  and  $n = 80$ . Finally, to study the effect of population size, we then repeated all the simulations based on the different values of  $\beta_1$ ,  $\beta_2$ ,  $m$ , and  $n$  when  $N = 500$  for three additional values of  $N$ , namely 5,000, 50,000, and 500,000. For each different combination of  $N$ ,  $\beta_1$ ,  $\beta_2$ ,  $m$ , and  $n$ , equations (4.8) through (4.10) were used to compute values for RV, EBU, EBJ, and RE analogous to those given in Table 1 for the estimator in (4.1). The results are provided in Table 2.

Generally speaking, the findings based on the results in Table 2 are similar to those arrived at for the estimators based on (4.1) and (4.4). In particular, the jackknife estimator of variance is more efficient than the usual estimator when the population size is sufficiently large. However, also of note is the fact that this efficiency seems to be related to the magnitude of the regression coefficients  $\beta_1$  and  $\beta_2$ ; that is, the jackknife estimator appears to achieve relatively greater efficiency for cases where the coefficient associated with the auxiliary variable  $X$ , is large relative to the analogous coefficient linked to  $Z$ .

**Table 2**  
Comparison of the jackknife and usual estimators of variance of the chain ratio estimator of the population mean where the auxiliary variable,  $X$ , follows a gamma distribution with a shape parameter of 2.2 and a scale parameter of one, and the auxiliary variable,  $Z$ , follows a gamma distribution with a shape parameter of 0.1 and a scale parameter of one

<i>m</i>	<i>n</i>	$\beta_1$	$\beta_2$	<i>N</i>	RV	EBU	EBJ	RE
100	20	3.5	2.5	500	0.769	0.000	0.027	1.063
				5,000	0.831	-0.012	0.020	2.282
				50,000	0.818	-0.006	0.028	1.785
				500,000	0.852	0.001	0.036	1.993
100	20	0.5	0.5	500	0.911	-0.001	0.004	0.791
				5,000	0.943	-0.001	0.002	0.888
				50,000	0.948	0.000	0.003	0.896
				500,000	0.946	0.000	0.003	0.899
100	20	3.5	0.5	500	0.845	-0.001	0.015	1.674
				5,000	0.932	-0.011	0.000	3.632
				50,000	0.947	-0.005	0.004	3.221
				500,000	0.947	0.000	0.010	3.637
100	20	0.5	2.5	500	0.866	-0.001	0.009	0.668
				5,000	0.858	-0.003	0.008	0.775
				50,000	0.855	-0.001	0.010	0.670
				500,000	0.855	0.000	0.012	0.697
400	80	3.5	2.5	500	0.540	0.000	0.013	0.044
				5,000	0.780	-0.001	0.009	1.346
				50,000	0.819	0.000	0.008	1.878
				500,000	0.810	-0.001	0.006	1.953
400	80	0.5	0.5	500	0.817	0.000	0.003	0.254
				5,000	0.956	0.000	0.000	0.885
				50,000	0.973	0.000	0.001	0.946
				500,000	0.973	0.000	0.000	0.963
400	80	3.5	0.5	500	0.579	0.000	0.010	0.041
				5,000	0.907	-0.001	0.003	3.158
				50,000	0.954	0.000	0.002	3.845
				500,000	0.950	-0.001	0.001	4.853
400	80	0.5	2.5	500	0.787	0.000	0.004	0.222
				5,000	0.862	0.000	0.002	0.570
				50,000	0.873	0.000	0.003	0.698
				500,000	0.875	0.000	0.002	0.595



Finally, an analogous simulation study was performed using the regression estimator of Ahmed (1997). However, the populations were created using  $\beta_1 = 10$  with  $\beta_2 = 0.5$ ,  $\beta_1 = 100$  with  $\beta_2 = 0.5$ ,  $\beta_1 = 0.5$  with  $\beta_2 = 10$ , and  $\beta_1 = 10$  with  $\beta_2 = 10$ . As before when the estimators of Rao and Sitter (1995), Sitter (1997), and Chand (1975) were considered, provided that the population is sufficiently large, the jackknife estimator of variance seems to be more efficient than the usual estimator.

## 5. Conclusion and discussion

In this paper, the problem of estimating the variance of various estimators of the population mean in two-phase sampling has been considered by jackknifing the famous two-phase calibrated weights of Hidiroglou and Särndal (1995, 1998). Simulation studies based on ratio, regression, and chain-type estimators suggest that provided that the population size is large enough and the first and second-phase samples are relatively small, the jackknife estimator of variance is more efficient than the usual estimator of variance, regardless of the estimator for the population mean that is considered. For small populations, it might be worthwhile to adjust the finite population correction factors instead of directly applying the jackknife technique. This is an area where further research could be conducted.

## Acknowledgements

This work was conducted while Sarjinder Singh was a postdoctoral fellow at Carleton University. The authors are grateful to the Associate Editor and the referees, whose comments greatly improved this manuscript. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

## Appendix

### Derivation of the jackknife estimator in (2.1)

In this part of the appendix, we prove (2.1) for the jackknifed estimator of the population mean in two phase-sampling. First, note that  $\hat{\beta}_1(j) = \hat{\beta}_1 + t_{1j} e_{1j}$  and  $\hat{\beta}_2(j) = \hat{\beta}_2 + t_{2j} e_{2j}$ , where  $t_{1j} = q_{1j} w_{1j}^o z_j / (q_{1j} w_{1j}^o z_j^2 - \sum_{i \in s_1} q_{1i} w_{1i}^o z_i^2)$ ,  $e_{1j} = x_j - \hat{\beta}_1 z_j$ ,  $t_{2j} = q_{2j} w_{2j}^o x_j / (q_{2j} w_{2j}^o x_j^2 - \sum_{i \in s_1} q_{2i} w_{2i}^o x_i^2)$ , and  $e_{2j} = y_j - \hat{\beta}_2 x_j$ . We also have  $\hat{Z}_1^o(j) = \hat{Z}_1^o + h_{1j}(\hat{Z}_1^o - z_j)$ ,  $\hat{X}_1^o(j) = \hat{X}_1^o + h_{1j}(\hat{X}_1^o - x_j)$ ,  $\hat{X}_2^o(j) = \hat{X}_2^o + h_{2j}(\hat{X}_2^o - x_j)$ , and  $\hat{Y}_2^o(j) = \hat{Y}_2^o + h_{2j}(\hat{Y}_2^o - y_j)$ , where  $h_{1j} = w_{1j}^o / (1 - w_{1j}^o)$  and  $h_{2j} = w_{2j}^o / (1 - w_{2j}^o)$ .

Using these results, for  $j \in s_2$ , we have

$$\begin{aligned} \hat{Y}^c(j) &= \hat{Y}_2^o + \hat{\beta}_2(\hat{X}_1^o - \hat{X}_2^o) + \hat{\beta}_1 \hat{\beta}_2(\bar{Z} - \hat{Z}_1^o) \\ &\quad + h_{2j}(\hat{Y} - y_j) + t_{2j} e_{2j}(\hat{X}_1^o - \hat{X}_2^o) \\ &\quad + \hat{\beta}_2 \{h_{1j}(\hat{X}_1^o - x_j) - h_{2j}(\hat{X}_2^o - x_j)\} \\ &\quad + t_{1j} e_{1j} \hat{\beta}_2(\bar{Z} - \hat{Z}_1^o) - t_{1j} e_{1j} \hat{\beta}_2 h_{1j}(\hat{Z}_1^o - z_j) \\ &\quad + \hat{\beta}_1 t_{2j} e_{2j}(\bar{Z} - \hat{Z}_1^o) - t_{2j} e_{2j} \hat{\beta}_1 h_{1j}(\hat{Z}_1^o - z_j) \\ &\quad - \hat{\beta}_1 \hat{\beta}_2 h_{1j}(\hat{Z}_1^o - z_j). \end{aligned}$$

Similarly, for  $j \in (s_1 - s_2)$ , we have

$$\begin{aligned} \hat{Y}^c(j) &= \hat{Y}_2^o + \hat{\beta}_2(\hat{X}_1^o - \hat{X}_2^o) + \hat{\beta}_1 \hat{\beta}_2(\bar{Z} - \hat{Z}_1^o) \\ &\quad + \hat{\beta}_2 h_{1j}(\hat{X}_1^o - x_j) + t_{1j} e_{1j} \hat{\beta}_2(\bar{Z} - \hat{Z}_1^o) \\ &\quad - t_{1j} e_{1j} \hat{\beta}_2 h_{1j}(\hat{Z}_1^o - z_j) \\ &\quad + \hat{\beta}_1 \hat{\beta}_2 \{(\bar{Z} - \hat{Z}_1^o) - h_{1j}(\hat{Z}_1^o - z_j)\}. \end{aligned}$$

Thus for  $j \in s_2$ ,

$$\begin{aligned} \hat{Y}^c(j) - \hat{Y}^c &= \{\hat{Y}_2^o(j) - \hat{Y}_2^o\} - \hat{\beta}_2(j) \{\hat{X}_2^o(j) - \hat{X}_2^o\} \\ &\quad - \hat{\beta}_1(j) \hat{\beta}_2(j) \{\hat{Z}_1^o(j) - \bar{Z}\} \\ &\quad + \hat{\beta}_2[\{\hat{X}_1^o(j) - \hat{X}_1^o\} - \hat{\beta}_1(j) \{\hat{Z}_1^o(j) - \bar{Z}\}] \\ &\quad + \hat{\beta}_2(j) \{\hat{X}_1^o(j) - \hat{X}_1^o\} \\ &\quad + \hat{\beta}_2[\{\hat{X}_2^o(j) - \hat{X}_1^o(j)\} \\ &\quad - \hat{\beta}_1(j) \{\bar{Z} - \hat{Z}_1^o(j)\} - \hat{\beta}_1 \{\bar{Z} - \hat{Z}_1^o\}], \end{aligned}$$

and for  $j \in (s_1 - s_2)$ ,

$$\hat{Y}^c(j) - \hat{Y}^c = \hat{\beta}_2[\{\hat{X}_1^o(j) - \hat{X}_1^o\} - \hat{\beta}_1(j) \{\hat{Z}_1^o(j) - \bar{Z}\}],$$

which proves (2.1).

### Consistency of the estimator of variance in (2.2)

In this part of the appendix, we prove that the estimator  $\hat{V}_{\text{JACK}}(\hat{Y}^c)$  in (2.2) is consistent. First, note that the variance of the estimator  $\hat{Y}^c$  defined in (1.6) can be approximated as:

$$\begin{aligned} V(\hat{Y}^c) &\approx V(\hat{Y}_2^o) + \beta_2^2[V(\hat{X}_1^o) + V(\hat{X}_2^o) - 2\text{Cov}(\hat{X}_1^o, \hat{X}_2^o)] \\ &\quad + \beta_1^2 \beta_2^2 V(\hat{Z}_1^o) \\ &\quad + 2\beta_2[\text{Cov}(\hat{Y}_2^o, \hat{X}_1^o) - \text{Cov}(\hat{Y}_2^o, \hat{X}_2^o)] \\ &\quad - 2\beta_1 \beta_2 \text{Cov}(\hat{Y}_2^o, \hat{Z}_1^o) \\ &\quad - 2\beta_1 \beta_2^2[\text{Cov}(\hat{X}_1^o, \hat{Z}_1^o) - \text{Cov}(\hat{X}_2^o, \hat{Z}_1^o)]. \end{aligned}$$

If it is assumed that  $\hat{\beta}_1(j) \approx \beta_1$ ,  $\hat{\beta}_2(j) \approx \beta_2$ , and similar to Rao and Sitter (1995), that  $\bar{x}_n(j)/\bar{x}_r(j) \approx \bar{x}_n/\bar{x}_r$ , it is quite straightforward to show that

$$\begin{aligned} \sum_{j \in S} [\hat{Y}^c(j) - \hat{Y}^c]^2 &\approx \sum_{j \in S_2} [\hat{Y}_2^o(j) - \hat{Y}_2^o]^2 + \hat{\beta}_2^2 \sum_{j \in S_2} [\hat{X}_2^o(j) - \hat{X}_2^o]^2 \\ &+ 2\hat{\beta}_2 \sum_{j \in S_2} [\hat{Y}_2^o(j) - \hat{Y}_2^o][\hat{X}_1^o(j) - \hat{X}_1^o] \\ &- 2\hat{\beta}_2 \sum_{j \in S_2} [\hat{Y}_2^o(j) - \hat{Y}_2^o][\hat{X}_2^o(j) - \hat{X}_2^o] \\ &- 2\hat{\beta}_2^2 \sum_{j \in S_2} [\hat{X}_1^o(j) - \hat{X}_1^o][\hat{X}_2^o(j) - \hat{X}_2^o] \\ &- 2\hat{\beta}_1 \hat{\beta}_2 \sum_{j \in S_2} [\hat{Y}_2^o(j) - \hat{Y}_2^o][\hat{Z}_1^o(j) - \hat{Z}_1^o] \\ &- 2\hat{\beta}_1 \hat{\beta}_2^2 \sum_{j \in S_2} [\hat{X}_2^o(j) - \hat{X}_2^o][\hat{Z}_1^o(j) - \hat{Z}_1^o] \\ &+ \hat{\beta}_2^2 \sum_{j \in S} [\hat{X}_1^o(j) - \hat{X}_1^o]^2 \\ &+ \hat{\beta}_1^2 \sum_{j \in S} [\hat{Z}_1^o(j) - \hat{Z}_1^o]^2 \\ &- 2\hat{\beta}_1 \hat{\beta}_2^2 \sum_{j \in S} [\hat{X}_1^o(j) - \hat{X}_1^o][\hat{Z}_1^o(j) - \hat{Z}_1^o]. \end{aligned}$$

Since the ten terms on the right hand side of this equation for  $\sum_{j \in S} [\hat{Y}^c(j) - \hat{Y}^c]^2$  are the consistent estimators of the analogous ten terms in the equation above for  $V(\hat{Y}^c)$ , it may be concluded that the jackknife estimator of variance in (2.2) is consistent.

## References

- Ahmed, M.S. (1997). The general class of chain estimators for the ratio of two means using double sampling. *Communications in Statistics, Theory and Methods*, 26(9), 2247-2254.
- Arnab, R., and Singh, S. (2006). A new method for estimating variance from data imputed with ratio method of imputation. *Statistics and Probability Letters*, 76, 513-519.
- Berger, Y. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 953-964.
- Berger, Y., and Skinner, C. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society, Series B*, 67, 79-89.
- Chand, L. (1975). *Some ratio type estimators based on two or more auxiliary variables*. PhD Thesis, Iowa State University, Ames, Iowa, USA.
- Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96, 260-269.
- Dewille, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 117-132.
- Hidiroglou, M.A., and Särndal, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Vol. II, 873-878.
- Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.
- Kim, J.K., Navarro, A. and Fuller, W.A. (2000). Variance estimation for 2000 Census coverage estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 515-520.
- Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.
- Kim, J.K., and Sitter, R.R. (2003). Efficient replication variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.
- Kott, P.S., and Stukel, D. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-89.
- Kovar, J., and Chen, E. (1994). Jackknife variance estimation of imputed survey data. *Survey Methodology*, 20, 45-52.
- Raj, D. (1965). On sampling over two occasions with probability proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.
- Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-60.
- Singh, S. (2000). Estimation of variance of regression estimator in two phase sampling. *Calcutta Statistical Association Bulletin*, 50, 49-63.
- Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.
- Strivenkataramana, T., and Tracy, D.S. (1989). Two-phase sampling for selection with probability proportional to size in sample surveys. *Biometrika*, 76, 818-821.

# A comparison of sample set restriction procedures

Jason C. Legg and Cindy L. Yu<sup>1</sup>

## Abstract

For many designs, there is a nonzero probability of selecting a sample that provides poor estimates for known quantities. Stratified random sampling reduces the set of such possible samples by fixing the sample size within each stratum. However, undesirable samples are still possible with stratification. Rejective sampling removes poor performing samples by only retaining a sample if specified functions of sample estimates are within a tolerance of known values. The resulting samples are often said to be balanced on the function of the variables used in the rejection procedure. We provide modifications to the rejection procedure of Fuller (2009a) that allow more flexibility on the rejection rules. Through simulation, we compare estimation properties of a rejective sampling procedure to those of cube sampling.

Key Words: Rejection sampling; Cube sampling; Stratification; Balanced sampling.

## 1. Introduction

A common practice in survey sampling is to utilize known population information about auxiliary variables to improve estimators of means and totals of characteristics of interest. When population control means or totals for an auxiliary variable are known, regression and other calibration estimators are often utilized. Let  $(\mathbf{x}_i, y_i, p_i)$ ,  $i = 1, 2, \dots, N$ , be a sequence of real vectors, where each  $\mathbf{x}_i$  is a  $k$  dimensional vector, and a sample  $A$  be selected from  $F_N = [(\mathbf{x}_1, y_1, p_1), \dots, (\mathbf{x}_N, y_N, p_N)]$  using a sample design with inclusion probabilities  $p_i$  and joint inclusion probabilities  $p_{ij}$ . Suppose the population mean of  $\mathbf{x}_i$ ,  $\bar{\mathbf{x}}_N$ , is known. Consider the regression estimator of the population mean of the form

$$\bar{y}_{\text{reg}} = \bar{\mathbf{z}}_N' \hat{\boldsymbol{\beta}}, \quad (1)$$

where  $\mathbf{z}_i$  contains design variables and  $\mathbf{x}_i$ ,  $\bar{\mathbf{x}}_N$  is the population mean of  $\mathbf{z}_i$ , and  $\hat{\boldsymbol{\beta}}$  is a regression coefficient estimator. For many designs,  $\hat{\boldsymbol{\beta}}$  of the form

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i \in A} \mathbf{z}_i \phi_i p_i^{-2} \mathbf{z}_i' \right)^{-1} \sum_{i \in A} \mathbf{z}_i \phi_i p_i^{-2} y_i, \quad (2)$$

where  $\phi_i$  are constants determined by the design, will be asymptotically efficient. Some examples of  $\phi_i$  choices are  $\phi_i = (1 - p_i)$  for Poisson sampling and for stratified random sampling,  $\phi_{hi} = (N_h - 1)^{-1} (N_h - n_h)$  for element  $i$  in stratum  $h$ . If we assume there is a vector  $\mathbf{d}$  such that

$$\phi_i p_i^{-2} \mathbf{z}_i' \mathbf{d} = p_i^{-1} \quad (3)$$

for all  $i$ , then estimator (1) is design consistent (Fuller 2002). The regression coefficient estimator (2) converges together with

$$\boldsymbol{\beta}_N = \left( \sum_{i=1}^N \mathbf{z}_i \phi_i p_i^{-1} \mathbf{z}_i' \right)^{-1} \sum_{i=1}^N \mathbf{z}_i \phi_i p_i^{-1} y_i.$$

As an example of applying equation (3), suppose we plan to select a Poisson sample and want to regress on a single covariate  $x_{1i}$  through the origin. If we add  $(1 - p_i)^{-1} p_i$  into  $\mathbf{z}_i$  to make  $\mathbf{z}_i' = (x_{1i}, [1 - p_i]^{-1} p_i)$ , then (1) will be design consistent for  $\bar{y}_N$  since (3) is satisfied by setting  $\mathbf{d}' = (0, 1)$ . If we further assume that a column of ones is in the column space of the regression variables  $\mathbf{z}_i$ , then for these  $\phi_i$  values, estimator (1) nearly attains the minimum asymptotic variance for design consistent regression estimators under certain regularity conditions (Rao 1994). An alternative approach to constructing a regression estimator is to start with a design consistent estimator, such as the generalized regression estimator of Särndal (1980), and determine the best coefficient given that form of the estimator. Starting with a design consistent form removes the need to satisfy (3). Condition (3) allows estimator (1) to be expressed in the form of a generalized regression estimator (Fuller 2009b, pages 116-117).

When auxiliary information is known at the unit level, the auxiliary information can also be incorporated into the sample design. For example in one classic case, the model with

$$y_i = \beta_0 + \beta_1 x_i + x_i \varepsilon_i, \quad (4)$$

$\varepsilon_i \sim \text{ind}(0, \sigma^2)$  and  $\text{cov}(\varepsilon_i, x_i) = 0$  is assumed for the population  $F_N$ . From Isaki and Fuller (1982), the optimal inclusion probabilities for the regression estimator are those that are proportional to the square root of the design variances, i.e.,  $p_i \propto x_i$  in this case. A possible sampling procedure is Poisson sampling with inclusion probabilities

1. Cindy L. Yu is an assistant professor in the Department of Statistics and the Center for Survey Statistics and Methodology at Iowa State University, Ames, IA 50010. E-mail: cindy.yu@iastate.edu; Jason C. Legg is a postdoctoral researcher at the Center for Survey Statistics and Methodology at Iowa State University, Ames, IA 50010. E-mail: jason-legg@hotmail.com.



$$p_i = \left| \sum_{i=1}^N x_i \right|^{-1} n_N x_i, \quad (5)$$

where  $n_N = \sum_{i=1}^N p_i$  is a specified target sample size. A second common design when model (4) is assumed is to stratify the population based on  $x$ . Strata are determined by setting the boundaries such that the sum of the sorted  $x_i$  values in each stratum are approximately equal. An equal number of units in each stratum are selected. This stratification design has the inclusion probabilities close to (5), and was shown to have an anticipated variance close to the best purposive sample model variance in the two-per-stratum case (Fuller 1981).

Another way to incorporate information from an auxiliary variable into the design is balancing. A sample  $A$  is balanced for variable  $z$  if

$$\bar{z}_{HT} = N^{-1} \sum_{i \in A} p_i^{-1} z_i = N^{-1} \sum_{i=1}^N z_i = \bar{z}_N. \quad (6)$$

A design is balanced for  $z$  if every sample with positive probability is balanced for  $z$ . Balancing can be thought of as calibration by design. To illustrate the effect of balancing, consider an equal inclusion probability design and  $z_i = (1, x_i)'$ . The conditional prediction variance of  $\bar{y}_{reg}$  under model (4) is

$$V(\bar{y}_{reg} - \bar{y}_N | \mathbf{x}, \bar{x}_{HT}) = E\{V(\bar{u}_{HT} | F_N) | \mathbf{x}, \bar{x}_{HT}\} + (\bar{x}_N - \bar{x}_{HT})^2 V(\hat{\beta}_1 | \mathbf{x}, \bar{x}_{HT}), \quad (7)$$

where  $u_i = x_i \varepsilon_i$ . For a balanced design, the second term in (7) is 0, which suggests we might improve the estimator by balancing on  $x$ . In practice, a combination of balancing and calibration will often outperform either technique used alone.

Balanced sample designs have some additional practical value. For many designs, there is a nonzero probability of selecting a sample that contains undesirable auxiliary variable values. For example, an undesirable sample could be a sample with insufficient sample allocation for domains or a sample with a large number of extreme values of auxiliary variables. Although stratified designs reduce the set of such possible samples by fixing the sample size within each stratum, undesirable samples could still be possible. For example, some stratified samples might have some negative weights from using regression estimators. Balancing can remove poor performing samples by only retaining samples with estimates close to known quantities and with only positive weights for regression estimators.

Balanced sampling was proposed by Royall and Cumberland (1981) as a way to reduce model bias from incorrectly specified polynomial superpopulation models. Valliant, Dorfman and Royall (2000) discuss the implications of balancing from a prediction approach to sampling.

Deville and Tillé (2004) investigated methods of selecting balanced samples within the design-based framework described above. See also Tillé (2006 Chapter 8) for a detailed treatment of balancing. In practice, finding a perfectly balanced design may not be possible. Very tight balancing can lead to a design with some extreme joint inclusion probabilities, including zero inclusion probabilities. Therefore, partial balancing is done in practice.

In this paper, we compare design properties through simulation studies of two balancing procedures, the rejective sampling of Fuller (2009a) and the cube sampling of Tillé (2006). We also provide modifications to Fuller's rejective sampling procedure that allow for more flexibility in balancing. In Section 2, the rejective sampling and the cube sampling are described. Properties of the inclusion probabilities of the two balancing procedures are compared in Section 3. In Section 4, some simulation results using balanced samples are presented. In Section 5, we provide adjustments to the rejective procedure. Concluding remarks are made in Section 6.

## 2. Balanced sampling procedures

Rejection sampling involves discarding any sample that does not meet a specified balancing tolerance. Fuller (2009a) presents one condition for rejecting a sample and Royall and Herson (1973) give another. In Fuller's procedure with the balancing variable vector  $z$ , a sample is selected under a specified initial design and retained if

$$(\bar{z}_{HT} - \bar{z}_N)' [V(\bar{z}_{HT} | F_N)]^{-1} (\bar{z}_{HT} - \bar{z}_N) < \gamma \quad (8)$$

for some constant  $\gamma > 0$ , where  $\bar{z}_{HT}$  is the Horvitz-Thompson mean estimator for variable  $z$ ,  $F_N$  is the given finite population,

$$V(\bar{z}_{HT} | F_N) = N^{-2} \sum_{i=1}^N \sum_{j=1}^N (p_{ij} - p_i p_j) z_i z_j' p_i^{-1} p_j^{-1},$$

$p_i$  is the inclusion probability for unit  $i$  and  $p_{ij}$  is the joint inclusion probability of unit  $i$  and unit  $j$  under the initial design. Otherwise, the sample is rejected, a new sample is selected under the initial design, and condition (8) is checked for the new sample. If the original design has a central limit theorem, the left side of (8) is asymptotically a  $\chi^2$  random variable with degrees of freedom equal to the number of auxiliary variables. An approximate rejection rate can be set using the quantiles of a  $\chi^2$  distribution for  $\gamma$ . Choice of a rejection rate will depend on objectives of each individual survey. Low rejection rates may not reduce the variance by a large amount, but provide sufficient comfort to a researcher that a very poor sample will not be selected. On the other hand, high rejection rates could provide large reductions in the variance, but the resulting samples could

have insufficient sample size to accommodate unplanned domain analysis. For example, if a researcher decides to conduct domain analysis on the tail of the distribution of a balancing variable, the joint inclusion probabilities could be small leading to few units in the domain for many samples.

The cube method was developed by Tillé and Deville and is described in Tillé (2006). The cube method attempts to select a balanced sample with predetermined first-order inclusion probabilities. If the first-order inclusion vector does not lead to a balanced design, an additional step of minimizing a cost constraint is used. Unlike the rejection procedure, higher order initial inclusion probabilities are not prespecified. The cost minimization step maintains the specified initial first-order inclusion probabilities.

As a way to understand the cube procedure, Tillé (2006) describes sampling geometrically. The set of all possible samples is defined to be the set of vectors for vertices of an  $N$  dimensional unit cube. For example, if  $N=3$ , the vertex  $(0, 1, 1)$  denotes a sample containing units two and three. Using the balancing equation (6) and desired  $p_i$  for  $i = 1, \dots, N$ , a balancing plane is created. Any sample where the balancing plane intersects a vertex of the unit  $N$  dimensional cube is a balanced sample. The design is balanced if every point of intersection between the balancing plane and the unit cube is a vertex of the unit cube. The cube sampling procedure begins by selecting a vector on the balancing plane, then a random walk from the initial point to an edge of the unit cube is done. Tillé refers to the random walk step as the flight phase. If the edge point at the end of the random walk is a vertex of the unit cube, the sample is selected. Otherwise, a cost minimization procedure is used to convert the fractional components of the edge vector to integers. The integer components of the edge vector are not changed in the cost minimization step. Tillé refers to the cost minimization step as the landing phase. Rejection sampling with high rejection rates produces results similar to cube sampling.

Other procedures besides rejection and cube sampling can be used to obtain nearly balanced samples. For example, stratification with boundaries determined by the  $x$  variables can also introduce some balancing effects to samples (Fuller 1981). Deciding the number of variables to use in the rejection and cube sampling procedures is essentially the same process as deciding how many variables to include in a regression estimator.

Software has been developed for selecting cube samples. For rejection sampling, standard software packages can be used to select a sample and compute (8). A loop needs to be written to complete the procedure. Programs for selecting cube samples have been written for SAS and R. See

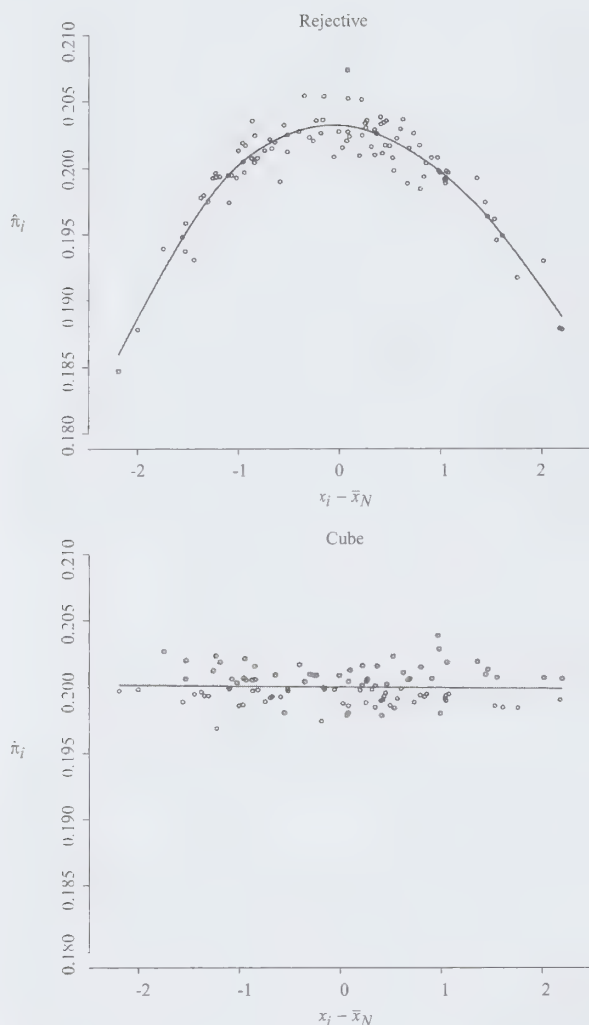
Rousseau and Tardieu (2004) for SAS and Matei and Tillé (2005) for R, and details of the procedures implemented are addressed in Deville and Tillé (2004). The R program available in the *sampling* library was used in the simulations in this paper. Because the cost minimization step of cube sampling is computationally intensive for more than 20 balancing variables, a variable suppression step is recommended for the landing phase in the programs.

### 3. Inclusion probabilities

Let  $\pi_i$  be the first-order inclusion probability for unit  $i$  and  $\pi_{ij}$  be the joint inclusion probability for unit  $i$  and  $j$  under a balanced design. Both rejective and cube sampling require initial first-order inclusion probabilities as inputs. The first-order inclusion probabilities are different than the initial values for rejection sampling. For rejection sampling, units closer to the population mean will have a slightly higher inclusion probability than units far from the mean. Cube sampling maintains the first-order inclusion probabilities from the initial specification. That is, for cube sampling  $\pi_i = p_i$ . Although for rejection sampling  $\pi_i \neq p_i$ , in general, the estimators considered will still use  $p_i$  rather than  $\pi_i$ .

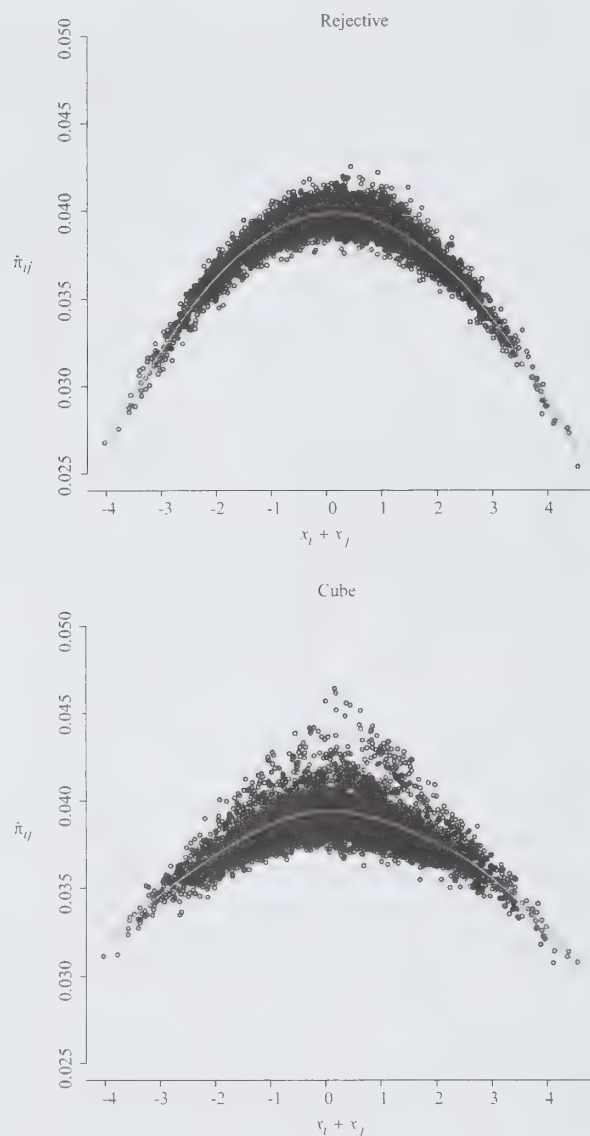
To illustrate differences between initial and final inclusion probabilities, samples of size 20 from a population of 100 units were simulated. The population of  $x$ -values was generated as random variables from a standard normal distribution. The rejection procedure used simple random sampling as the initial design and balanced on  $x$ . The cube sample procedure used a balancing vector of  $\mathbf{z}_i = (p_i, x_i)'$ , where  $p_i = 20/100$  for all  $i$ . The inclusion of  $p_i$  in the balancing vector for cube sampling was to control the sample size so that the resulting design would be comparable to using an initial design of simple random sample design in the rejection sampling simulation. First-order inclusion probabilities were estimated using a Monte Carlo simulation of size 100,000 (Figure 1). The curve was obtained by nonparametric fitting. An approximate 90% rejection rate was used for the rejection sampling. From rejection sampling theory, first-order inclusion probabilities are approximately a quadratic function of the distance  $x_i - \bar{x}_N$  for an equal probability initial sample design (Fuller 2009a). The plot suggests that all first-order inclusion probabilities are 0.2 for the cube sample design. As expected, Figure 1 indicates the cube method maintains the specified first-order inclusion probabilities, but the rejective does not. As a result, the Horvitz-Thompson estimator using the initial inclusion probabilities ( $p_i$ ) and the rejective samples is biased.





**Figure 1** Simulated first-order inclusion probabilities. The balancing variable for the rejective method is  $z_i = x_i$ , and for the cube method is  $z_i = (p_i, x_i)'$ , where  $p_i = 20/100$

The joint inclusion probabilities for the rejection sampling procedure differ from those of the initial design. A pair of units  $i$  and  $j$  are likely to have a high joint inclusion probability if  $x_i + x_j - 2\bar{x}_N$  is close to zero for an equal probability initial sample design. The joint inclusion probabilities were estimated from simulated samples of size 20 from 100 (Figure 2). The joint inclusion probability for simple random sampling is 0.038. The rejection sampling joint inclusion probabilities are approximately a quadratic function of  $x_i + x_j$ . The plot of cube sampling joint inclusion probabilities against  $x_i + x_j$  appears to have sharper angles than the rejection joint inclusion probabilities. High joint inclusion probabilities for the cube design are associated with pairs of units that are on the far opposite sides of  $\bar{x}_N$ . That is, for the sample value of  $x_i + x_j$ , those pairs with a large value of  $|x_i| + |x_j|$  have a large probability of inclusion (Figure 3).



**Figure 2** Simulated second-order inclusion probabilities. The balancing variable for the rejective method is  $z_i = x_i$ , and for the cube method is  $z_i = (p_i, x_i)'$ , where  $p_i = 20/100$

The Horvitz-Thompson estimator using the initial inclusion probabilities under rejection sampling has an  $O_p(n^{-1})$  bias while the Horvitz-Thompson estimator under cube sampling is unbiased. The standard Horvitz-Thompson variance estimator is biased for both procedures. Using Monte Carlo methods, the inclusion probabilities can be estimated so that nearly unbiased Horvitz-Thompson estimators can be used. However, for a large population, simulating enough samples to give a precise estimate of the joint inclusion probability for each pair of units is impractical. An alternative approach to variance estimation is to use a regression estimator and the variance estimator for the regression estimator. This is intuitively appealing because balancing is similar to regression through design.



Upon using the regression estimator, the bias of the regression estimator under both cube and rejective methods is of the same order. For rejective sampling, Fuller (2009a) gives conditions for the consistency of the variance estimator for the regression estimator. For cube sampling, Deville and Tillé (2005) and Tillé (2006) suggest using the variance estimator for a regression estimator furnishes a good approximation to the variance of the Horvitz-Thompson estimator. The variance estimators proposed by Deville and Tillé (2005) perform well when the joint inclusion probabilities of the resulting cube design are approximately equal to joint inclusion probabilities from a Poisson design. In the simulation studies of Section 4, the variance estimators proposed in Fuller (2009a) and Deville and Tillé (2005) are evaluated.

#### 4. Simulation of the regression estimator

A population of size 100 was generated from the model

$$y_i = x_i + 0.55x_i^2 + x_i\varepsilon_i \quad (9)$$

$\varepsilon_i \sim \text{iid } N(0, 0.4)$ , where the  $x_i$  are fixed values in the range of 0 to 4 (Figure 4). Seventy-two of the  $x$  values were randomly simulated values less than 1.15 from a standard exponential distribution. The remaining 28 values, ranging from 0.18 to 4.0, were deterministically added to form the data set of  $x$ . The fixed  $x$  values were selected to be fairly right skewed so that some large and small strata when stratifying the population on  $x$  with approximately equal within-stratum sum of sorted  $x_i$  will be produced. The population was held fixed after initial selection. Model (9) contains a quadratic term, and was picked to simulate performance of the design and estimator strategy when model (4) was assumed in design and estimation.

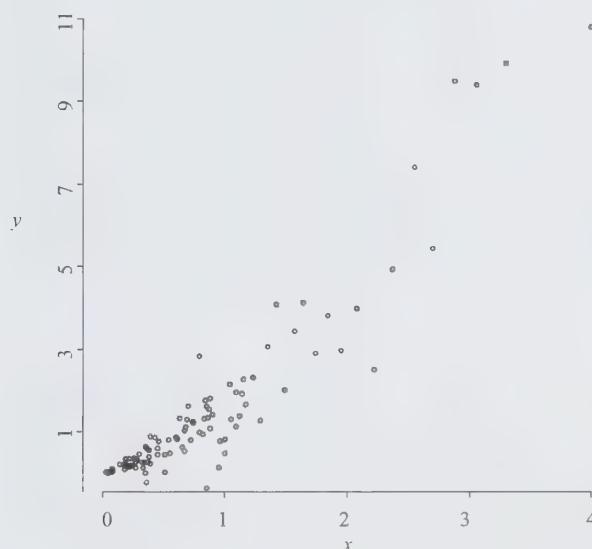


Figure 4 Simulation population under model (9)

We consider Poisson sampling and two-per-stratum stratified random sampling as initial designs. Strata were determined by setting the boundary so that the within stratum sum of sorted  $x_i$  was roughly equal for all strata. The sample size was set to 20, and ten strata were formed. The stratum sizes were 35, 15, 11, 9, 8, 7, 5, 4, 3, and 3. The rejection procedure used a stratified two-per stratum sample selection with equal inclusion probabilities within a stratum. The stratum boundaries were chosen this way so that the inclusion probability of unit  $i$  is closely proportional to  $x_i$ , which is the optimal inclusion probability under model (9) (Ikasi and Fuller 1982). Such a stratified design can also be partially balanced on  $x$  through a standard design. Balance

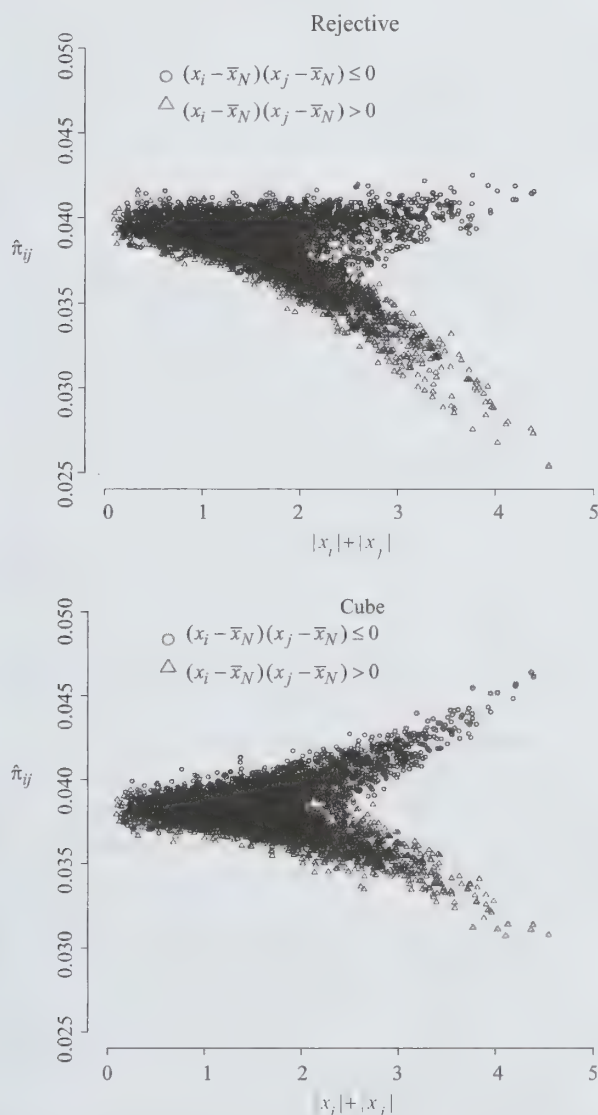


Figure 3 Simulated second-order inclusion probabilities with absolute sums of  $x$ . The balancing variable for the rejective method is  $z_i = x_i$ , and for the cube method is  $z_i = (p_i, x_i)'$ , where  $p_i = 20/100$

in the stratified random sampling design is achieved using a step function to approximate a line. The stratified design will also be partially balanced on  $x^2$ . The stratified random sample design is intended to illustrate how much more one can benefit from additional balancing. Two units per stratum were drawn in order to obtain the maximum number of strata while still permitting unbiased variance estimation. Fuller (1981) showed that, in the two-per-stratum case, this stratified design has an anticipated variance close to the best purposive model variance under (4). Initial inclusion probabilities for the Poisson design with expected sample size 20 were set to the initial inclusion probabilities of the stratified design.

The regression estimator considered in this paper is in the form of (1) with  $\hat{\beta}$  defined in (2). The regression variable  $\mathbf{z}$  is a vector of auxiliary variables that contains design variables and  $x$ . For the Poisson designs, we used  $\mathbf{z}_i = (1, p_i, x_i, (1 - p_i)^{-1} p_i)'$  as the vector of balancing variables and as the regression variable vector. The first variable provides control for population size, the second variable is a control for sample size, the third variable provides balance on  $x$ , and the fourth variable guarantees that the regression estimator is design consistent. See condition (3) for the design consistency of  $\bar{y}_{\text{reg}}$  and set  $\mathbf{d} = (0, 0, 0, 1)'$ . For two-per-stratum stratified samples, the vector of balancing variables is  $(x_i, I_{1i}, I_{2i}, \dots, I_{10i})'$  for cube sampling, where  $I_{hi}$  are the stratum indicator variables defined as

$$I_{hi} = \begin{cases} 1 & \text{unit } i \text{ in stratum } h \\ 0 & \text{otherwise} \end{cases}$$

for  $h = 1, 2, \dots, 10$ . Only the  $x$  variable is included in the rejective balancing procedure since the sample from this initial design is automatically balanced on the stratum indicator variables. The regression variable vector for both balancing procedures is  $\mathbf{z}_i = (x_i, I_{1i}, \dots, I_{10i})'$ .

For the initial designs, the variance estimators for  $\bar{y}_{\text{reg}}$  are the variance estimators of the mean of  $e_i = y_i - \mathbf{z}_i' \beta_N$  calculated with  $\hat{e}_i$ , where  $\hat{e}_i = y_i - \mathbf{z}_i' \hat{\beta}$ . For Poisson sampling, the variance estimator is

$$\hat{V}(\bar{y}_{\text{reg}}) = (n - s)^{-1} n \bar{\mathbf{z}}_N' \hat{\mathbf{M}}_{zz}^{-1} \sum_{i \in A} \mathbf{z}_i p_i^{-4} \times (1 - p_i)^3 \hat{e}_i^2 \mathbf{z}_i' \hat{\mathbf{M}}_{zz}^{-1} \bar{\mathbf{z}}_N, \quad (10)$$

where

$$\hat{\mathbf{M}}_{zz} = N^{-1} \sum_{i \in A} \mathbf{z}_i p_i^{-2} (1 - p_i) \mathbf{z}_i',$$

and  $s$  is the number of variables in  $\mathbf{z}$ . Derivation of (10) is provided in the appendix.

For stratified random sampling with two-per-stratum, the variance estimator for  $\bar{y}_{\text{reg}}$  is

$$\begin{aligned} \hat{V}(\bar{y}_{\text{reg}}) &= (H - 1)^{-1} H \sum_{h=1}^H [(1 - W_h)^{1/2} \\ &\quad \{0.5 W_h + (\bar{\mathbf{z}}_N - \bar{\mathbf{z}}) \hat{\mathbf{M}}_{zz, h}^{-1} \phi_h W_h^2 (\mathbf{z}_{h1} - \mathbf{z}_{h2})\} \\ &\quad \times (\hat{e}_{h1} - \hat{e}_{h2})]^2, \end{aligned} \quad (11)$$

where

$$\hat{\mathbf{M}}_{zz, h} = N_h^{-1} \sum_{i \in A_h} \mathbf{z}_i p_i^{-2} \phi_h \mathbf{z}_i',$$

$A_h$  is the sample set in stratum  $h$ ,  $W_h = n_h/N_h$ ,  $\phi_h = (N_h - 1)^{-1}(N_h - 2)$  for units in stratum  $h$ ,  $\mathbf{z}_{hi}$  is the auxiliary variable vector  $\mathbf{z}_i$  in stratum  $h$ ,

$$\hat{e}_{hi} = y_{hi} - \bar{y}_h - (\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)' \hat{\beta},$$

$\bar{y}_h$  and  $\bar{\mathbf{z}}_h$  are stratum means of  $y_{hi}$  and  $\mathbf{z}_{hi}$ , respectively, and  $H = 10$  is the number of strata. The derivation of (11) follows the same approach to the one in appendix and has been omitted.

For rejective sampling, the same variance estimators (10) and (11) using the initial design inclusion probabilities, were used to compute the variance estimator of  $\bar{y}_{\text{reg}}$  for rejective samples. Fuller (2009a) proved that the large sample properties of the regression estimator for the rejective sample are the same as those of the regression estimator for the original inclusion procedure under some regularity conditions. For cube sampling, a variance estimator proposed by Deville and Tillé (2005) was evaluated for  $\bar{y}_{\text{reg}}$  using cube samples.

Let  $p(\cdot)$  denote the initial design and  $\pi(\cdot)$  be the resulting scheme after balancing. The number of samples selected was 30,000 for each Monte Carlo simulation under initial designs, cube sampling and rejective sampling with both 90% and 95% rejection rates. The Horvitz-Thompson estimator  $\bar{y}_{\text{HT}}$  and the regression estimator  $\bar{y}_{\text{reg}}$  were constructed using initial inclusion probabilities  $p_i$ . Note that for rejection sampling, the Horvitz-Thompson estimator using the initial inclusion probabilities is not the Horvitz-Thompson estimator under the balanced designs. For each initial design, the following quantities were computed in the simulation studies.

- $V_p(\bar{y}_{\text{HT}})$  (or  $V_p(\bar{y}_{\text{reg}})$ ): Monte Carlo variance of the Horvitz-Thompson estimator (or the regression estimator) using samples from initial designs.
- $V_\pi(\bar{y}_{\text{HT}})$  (or  $V_\pi(\bar{y}_{\text{reg}})$ ): Monte Carlo variance of the Horvitz-Thompson estimator (or the regression estimator) for balanced samples.
- $\text{bias}_\pi(\bar{y}_{\text{HT}})$  (or  $\text{bias}_\pi(\bar{y}_{\text{reg}})$ ): Monte Carlo bias of the Horvitz-Thompson estimator (or the regression estimator) using balanced samples.



For cube samples,

- $\hat{V}_{DT}(\bar{y}_{reg})$ : estimated variance of the regression estimator using the variance estimators in Deville and Tillé (2005) and each cube sample.
- $\text{ave}(\hat{V}_{DT}(\bar{y}_{reg}))$ : Monte Carlo average of  $\hat{V}_{DT}(\bar{y}_{reg})$  using all cube samples.

Deville and Tillé (2005) recommend several variance estimators based on a Poisson sampling approximation with corrections for known constraints in the design variance. The first three estimators in Deville and Tillé (2005) have minor differences, therefore only the second estimator was used in the simulation studies. Deville and Tillé (2005) also propose the fourth estimator, but that estimator requires solving a nonlinear equation system, which would have been computationally expensive to add to the simulation. However, the fourth estimator could perform better than the other cases for stratified designs, since their fourth estimator reproduces the variance of a stratified random sample when the balancing vector contains stratum indicators.

For rejective samples,

- $\hat{V}(\bar{y}_{reg})$ : estimated variance of the regression estimator using equation (10) (or (11)) for the Poisson (or two-per-stratum stratified) initial design and each balanced sample.
- $\text{ave}(\hat{V}(\bar{y}_{reg}))$ : Monte Carlo average of  $\hat{V}(\bar{y}_{reg})$  using all balanced samples.

In the simulations,  $\hat{V}(\bar{y}_{reg})$  was also computed for cube samples, for comparison.

Table 1 reports the estimates for the Poisson design. The variance of the Horvitz-Thompson mean under initial Poisson sampling with expected sample size 20 and no balancing is  $V_p(\bar{y}_{HT}) = 0.08$ . The variances in Table 1 are standardized by  $V_p(\bar{y}_{HT})$ , and the biases are standardized by  $\sqrt{V_p(\bar{y}_{HT})}$ . The Horvitz-Thompson estimator is unbiased under the cube method designs, because cube sampling retains the first order inclusion probabilities. The Horvitz-Thompson estimator using initial design inclusion probabilities is biased under rejective sampling since the inclusion probabilities differ from the initial design inclusion probabilities, as indicated in Figure 1. The bias of the regression estimator under rejective sampling is less than the bias of the Horvitz-Thompson estimator with initial design inclusion probabilities. The bias of  $\bar{y}_{reg}$  under both cube and rejective procedures is of the same order. Increasing the rejection rate increases the bias of  $\bar{y}_{reg}$  for the rejection designs. However, the biases in  $\bar{y}_{reg}$  under both balancing procedures and rejection rates are negligible relative to the Monte Carlo variances. For the Horvitz-Thompson estimator using initial design inclusion probabilities, the gain from using the balanced sample is substantial for both cube

and rejective methods. The mean squared errors are further reduced by using the regression estimator along with either balancing procedures. The gain from using the regression estimator is larger for rejective sampling than for cube sampling, likely due to the cube method achieving tighter balance than the rejective method. Both procedures lead to similar variances for the regression estimator. The variance of the regression estimator under the Poisson initial design is  $V_p(\bar{y}_{reg}) = 0.249$  (relative to  $V_p(\bar{y}_{HT})$ ). By comparing 0.249 to the fourth row of Table 1, we can see that the gain from using the balanced samples on the regression estimator is moderate. The result is consistent with the finding in Fuller (2009a) that the variance reduction in  $\bar{y}_{reg}$  by using rejective samples is due to a second order correction. The variance estimator of  $\bar{y}_{reg}$  using (10) has small bias for both cube and rejective samples ( $\text{ave}(\hat{V}(\bar{y}_{reg}))$  in Table 1). The variance estimator  $\hat{V}_{DT}(\bar{y}_{reg})$  proposed in Deville and Tillé (2005) performed similarly as  $\hat{V}(\bar{y}_{reg})$  in (10) since the second variance estimator in Deville and Tillé (2005) is very close to (10) for Poisson sampling. This result supports the claim that the Poisson approximation assumption in the variance estimators of Deville and Tillé (2005) is satisfied for the Poisson design case.

**Table 1**  
Properties of samples based on Poisson sampling of expected size 20.  $V_p(\bar{y}_{HT}) = 0.08$  and  $V_p(\bar{y}_{reg})/V_p(\bar{y}_{HT}) = 0.249$

	Cube	Rej. 90%	Rej. 95%
$\text{bias}_{\pi}(\bar{y}_{HT})/\sqrt{V_p(\bar{y}_{HT})}$	-0.002	-0.016	-0.007
$\text{bias}_{\pi}(\bar{y}_{reg})/\sqrt{V_p(\bar{y}_{HT})}$	-0.002	0.002	0.005
$V_{\pi}(\bar{y}_{HT})/V_p(\bar{y}_{HT})$	0.142	0.270	0.220
$V_{\pi}(\bar{y}_{reg})/V_p(\bar{y}_{HT})$	0.131	0.136	0.129
$\text{ave}(\hat{V}(\bar{y}_{reg}))/V_p(\bar{y}_{HT})$	0.122	0.123	0.121
$\text{ave}(\hat{V}_{DT}(\bar{y}_{reg}))/V_p(\bar{y}_{HT})$	0.120	-	-

In Table 2, estimates under the initial two-per-stratum stratification design are reported. The variance of the Horvitz-Thompson mean under the initial stratification design is  $V_p(\bar{y}_{HT}) = 0.011$  and all estimates are standardized by this value. Since stratification in this initial design controls for most of the effect of  $x$  on  $y$ , the regression estimator is not a large improvement over the Horvitz-Thompson estimator using initial design inclusion probabilities. The bias and variance of  $\bar{y}_{HT}$  are close to those of  $\bar{y}_{reg}$  under both cube and rejective methods. The larger estimated bias in  $\bar{y}_{HT}$  under cube sampling is due to Monte Carlo error. The gain from balancing on  $x$  is not large, compared to the gain in the Poisson example. However, with this highly controlled initial stratified design, in which



the initial samples are already partially balanced on  $x$ , there still can be a modest benefit from additional balancing and using  $\bar{y}_{\text{reg}}$  estimators. This result is seen for  $\bar{y}_{\text{reg}}$  by comparing the fourth row of Table 2 to the variance of  $\bar{y}_{\text{reg}}$  under the initial design  $V_p(\bar{y}_{\text{reg}}) = 0.987$ . Therefore, in this case a good strategy is to combine stratification, balancing, and regression, which is a similar conclusion drawn in Deville and Tillé (2004). The variance estimator  $\hat{V}(\bar{y}_{\text{reg}})$  using (11) gives estimates on average for the regression estimator variances under both cube and rejective procedures that are close to the true variances. However, the variance estimator  $\hat{V}_{DT}(\bar{y}_{\text{reg}})$  proposed by Deville and Tillé (2005) performed poorly for cube sampling. A possible reason is that the Poisson sampling approximation in the second variance estimator of Deville and Tillé (2005) assumes joint inclusion probabilities that are far from the actual joint inclusion probabilities in the small strata. The joint inclusion probabilities in the small strata are closer to those of stratified random sampling than Poisson sampling. This issue might explain why  $\hat{V}(\bar{y}_{\text{reg}})$  in (11) using the initial two-per-stratum inclusion probabilities is less biased than  $\hat{V}_{DT}(\bar{y}_{\text{reg}})$  in this case.

**Table 2**  
Properties of samples based on stratified sampling of size 20.  
 $V_p(\bar{y}_{\text{HT}}) = 0.011$  and  $V_p(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}}) = 0.987$

	Cube	Rej. 90%	Rej. 95%
$\text{bias}_{\pi}(\bar{y}_{\text{HT}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0.028	0.014	0.010
$\text{bias}_{\pi}(\bar{y}_{\text{reg}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0.013	0.014	0.010
$V_{\pi}(\bar{y}_{\text{HT}})/V_p(\bar{y}_{\text{HT}})$	0.910	0.866	0.813
$V_{\pi}(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}})$	0.929	0.865	0.813
$\text{ave}(\hat{V}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0.907	0.881	0.775
$\text{ave}(\hat{V}_{DT}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0.792	-	-

To assess large sample properties of the balancing procedures, the size of the Poisson simulation was quadrupled. The population was replicated four times and a sample of expected size 80 was selected. The Horvitz-Thompson variance of a mean under the Poisson design is  $V_p(\bar{y}_{\text{HT}}) = 0.020$  and the regression estimator variance is  $V_p(\bar{y}_{\text{reg}}) = 0.132$ . The resulting relative variances and biases are close to the results for samples of size 20 (Table 3). The simulation results agree with the theoretical result of Fuller (2009a) that the regression estimator is an  $O_p(n^{-1/2})$  estimator after rejection of the type used in this paper. Although it has not been proven here, regression estimator after cube sampling appears to possess similar properties to the regression estimator using rejection sampling.

**Table 3**  
Properties of samples based on Poisson sampling of expected size 80.  $V_p(\bar{y}_{\text{HT}}) = 0.02$  and  $V_p(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}}) = 0.132$

	Cube	Rej. 90%	Rej. 95%
$\text{bias}_{\pi}(\bar{y}_{\text{HT}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	0.002	-0.006	-0.007
$\text{bias}_{\pi}(\bar{y}_{\text{reg}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	0.002	0.000	-0.001
$V_{\pi}(\bar{y}_{\text{HT}})/V_p(\bar{y}_{\text{HT}})$	0.127	0.267	0.224
$V_{\pi}(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}})$	0.122	0.124	0.123
$\text{ave}(\hat{V}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0.121	0.121	0.121
$\text{ave}(\hat{V}_{DT}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0.121	-	-

## 5. Adjustments to the rejection procedure

Fuller's rejection sampling procedure treats all balancing variables with the same importance. For a large number of balancing variables, exact balance on all variables cannot be expected and the approximation could be poor for some important variables. Therefore, a practitioner may want to have tighter balance on a subset of the balancing variables. As an example, a researcher may want to use Poisson sampling for simplicity but also have some control on the random sample size. A random sample size can complicate study planning and is a large contributor to the variance of estimators. Balanced sampling can be used to reduce the variation in sample sizes by balancing on the variable  $p_i$ , which is the initial first-order inclusion probability. For Fuller's rejection procedure, the variance of the sample size increases when the number of balancing variables increases and the rejection rate is held constant. The rejection procedure can be altered so that the  $p_i$  balance is tighter than the balance for other variables.

One approach to increasing the balancing on a subset of variables is to change the rejection test function. The order of the approximation to the first and second-order inclusion probabilities in Fuller (2009a) remains the same when the variance matrix in the rejection quadratic form is replaced with a symmetric positive definite matrix of the same order.

To determine weights for weighted rejection sampling, it is convenient to transform the balancing variables so that  $V(\bar{z}_{\text{HT}} | F_N)$  is a diagonal matrix. The weighted rejection sampling test statistic is

$$\sum_{q=1}^m c_q V(\bar{z}_{\text{HT}, q} | F_N)^{-1} (\bar{z}_{\text{HT}, q} - \bar{z}_{N, q})^2, \quad (12)$$

where  $m$  is the number of balancing variables,  $z_q$  is the  $q^{\text{th}}$  balancing variable, and  $c_q$  are selected weights. The weight on the first variable  $z_{1i} = p_i$  can be set large relative to the weights on other variables to reduce variation in sample size. The transformation is the Gramm-Schmidt transformation using the design variances under the initial design. Balancing

is done on the transformed variables, but the first variable is not transformed. The transformed variables have uncorrelated Horvitz-Thompson estimators. Balancing on the transformed variables will still balance the original variables since each transformed variable is a residual from a regression operation on preceding variables.

Equation (12) can be paralleled to the penalty term of the distance function underlying ridge calibration. See Rao and Singh (1997), Beaumont and Bocci (2008), and Chambers (1996). Specifically, selection of the  $c_q$  weights is similar to the problem of selecting appropriate costs in ridge calibration. Thus, rejection sampling using (12) can be viewed as incorporating ridge calibration at the design stage.

A second way to produce tighter balance on a subset of variables is to do rejection separately for subsets. A test statistic is produced for each subset and a sample must be accepted by all of the tests to be accepted. In the Poisson case, one test statistic may reject if the sample size is not within a specified tolerance of the expected sample size. This second approach requires some additional assumptions beyond those in Fuller (2009a), but a similar argument can be used to justify the procedure.

To prove the convergence properties of the multiple test rejection procedure, it is convenient to consider two subsets of balancing variables and think of rejection being done sequentially on each subset. We call the two subset rejection procedure a two-step rejective sampling procedure. Suppose  $\mathbf{z}'_i = (z'_{i1}, z'_{i2})$  is the balancing vector and the original design is denoted as  $p(\cdot)$ . The procedure is as follows.

Step 1: Select a sample using  $p(\cdot)$  and reject samples with the balancing condition (8) on the first subset  $\mathbf{z}_1$ ,

$$Q_1 = (\bar{\mathbf{z}}_{HT,1} - \bar{\mathbf{z}}_{N,1})' V(\bar{\mathbf{z}}_{HT,1} | F_N)^{-1} (\bar{\mathbf{z}}_{HT,1} - \bar{\mathbf{z}}_{N,1}) < \gamma_1.$$

Step 2: Use the accepted sample from step 1 to check the balancing condition (8) on the second subset  $\mathbf{z}_2$ ,

$$Q_2 = (\bar{\mathbf{z}}_{HT,2} - \bar{\mathbf{z}}_{N,2})' V(\bar{\mathbf{z}}_{HT,2} | F_N)^{-1} (\bar{\mathbf{z}}_{HT,2} - \bar{\mathbf{z}}_{N,2}) < \gamma_2.$$

Reject the sample if the condition is not satisfied and repeat Step 1.

In both weighted and two-step procedures, trial and error is likely needed to choose  $\gamma$ 's in practice. In the weighted procedure, the quadratic form becomes a sum of multiples of  $\chi^2$  random variables, which makes selection of  $\gamma$  more difficult than in the unweighted case. We used moment matching approximations to select  $\gamma$ 's that provide rejection rates close to desired, but then resorted to small simulations to determine the rejection rate as a function of  $\gamma$ . For the two-step procedure, we used a  $\chi^2$  approximation to select a  $\gamma_1$  that gave approximately the desired rejection rate at the first step, and used second  $\chi^2$  approximation to select an initial  $\gamma_2$  that gave approximately the desired

rejection rate at the second step. The second parameter  $\gamma_2$  was adjusted in order to achieve the target overall rejection rate. The choice of  $\gamma$ 's in the two-step procedure is subjective because many combinations of  $\gamma_1$  and  $\gamma_2$  can produce the same overall rate. In practice, a practitioner likely will set a tight bound for the first variable subset and loose bounds on the remaining balancing variables.

The large sample mean and variance of the regression estimator under the two-step rejective sample are the same as those of the regression estimator for the original design. Also, the usual estimator of variance under the original design for the regression estimator is appropriate for the two-step rejective sample. The proof of this statement is an extension of the proof in Fuller (2009a) and can be provided upon request.

To examine some properties of the two procedures, the Monte Carlo simulations for the Poisson initial sample design were repeated with the variable  $p_i$  separated from the other three variables. The balancing vector was transformed so that the variance matrix of the Horvitz-Thompson total estimators was diagonal. For the weighting procedure, the weight on the  $p_i$  component of the quadratic form was set to 1.5, the weights on the other components were set to 1, and  $\gamma$  was set to 0.627. This weighting procedure restricted the samples to those with sample sizes ranging from 18 to 22. For the two-step procedure, any sample with a sample size outside of the range from 18 to 22 was rejected in the first step and then the quadratic form for the remaining three variables was checked using a  $\gamma$  of 0.63 for the second step. Given the good performance of the variance estimator  $\hat{V}(\bar{y}_{reg})$  in (10), Table 4 only contains its Monte Carlo averages values  $\text{ave}(\hat{V}(\bar{y}_{reg}))$ .

**Table 4**  
Properties of rejection samples with adjustments based on Poisson sampling of expected size 20, and 95% rejection rate

	Weighted	Two-step
$\text{bias}_\pi(\bar{y}_{HT}) / \sqrt{V_p(\bar{y}_{HT})}$	-0.005	-0.014
$\text{bias}_\pi(\bar{y}_{reg}) / \sqrt{V_p(\bar{y}_{HT})}$	0.003	0.002
$V_\pi(\bar{y}_{HT}) / V_p(\bar{y}_{HT})$	0.210	0.217
$V_\pi(\bar{y}_{reg}) / V_p(\bar{y}_{HT})$	0.132	0.132
$\text{ave}(\hat{V}(\bar{y}_{reg})) / V_p(\bar{y}_{HT})$	0.121	0.121
$V_\pi(n)$	1.237	1.902

Results for expected sample size of 20 and a rejection rate near 95% were similar for the two adjustment procedures (Table 4). The Horvitz-Thompson estimator for the weighted procedure performed slightly better than the Horvitz-Thompson estimator for the two-step procedure. A reason for this discrepancy is that the weighted procedure

had much less variation in sample sizes ( $V_\pi(n)$  in the last row of Table 4). Additional simulations with larger expected sample sizes gave similar relative variances. The regression estimator performed at roughly the same efficiency for the two procedures. The Horvitz-Thompson estimators using the initial design inclusion probabilities for these adjustment procedure performed slightly better than the Horvitz-Thompson estimator for the rejection procedure that did not place additional control on the sample size.

## 6. Discussion

Rejection sampling and cube sampling produce roughly equally performing regression estimators. Balancing provides major gains when the initial design provides little control on the auxiliary values entering samples. A well stratified sample design provides many of the benefits of balancing on a continuous variable. However, further balancing after stratification can still yield small mean squared error gains for regression estimators. Additionally, balancing could be used to prevent negative weights produced by regression estimators (Fuller 2009a).

For the simulations, the rejection rate was fixed at 90% for the larger population. When the population and sample sizes are increased, the rejection rate can be increased while still maintaining a large set of possible samples. Additional simulations were carried out with rejection rates near 99%, but the results were not presented since the differences between the results with 95% and with 99% were very small and the bias of  $\bar{y}_{\text{reg}}$  remained negligible. The marginal variance reduction due to balancing decreases as the balancing condition is tightened.

In some special cases, an investigator may want to balance tightly on some variables and weakly on others. Gains can be made by choosing different weights for different variables or by dividing the variables into separate test sets. The weighted and two-step rejection procedures performed comparably, so the decision between procedures will largely be based on the ease of implementation.

## Acknowledgements

This work was supported by Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University. The authors thank Wayne A. Fuller, the associate editor and two anonymous referees for helpful comments that improve the paper.

## Appendix

Start with

$$V(\bar{y}_{\text{reg}} | F_N) = V(\bar{y}_{\text{reg}} - \bar{y}_N | F_N).$$

Let

$$\bar{y}_N = \bar{\mathbf{z}}_N' \boldsymbol{\beta}_N$$

and note

$$y_i = \mathbf{z}_i' \boldsymbol{\beta}_N + e_{Ni}$$

$$\hat{\boldsymbol{\beta}} = \left[ \sum_{i \in A} \mathbf{z}_i \phi_i p_i^{-2} \mathbf{z}_i' \right]^{-1} \sum_{i \in A} \mathbf{z}_i \phi_i p_i^{-2} (\mathbf{z}_i' \boldsymbol{\beta}_N + e_i)$$

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_N + \left[ N^{-1} \sum_{i \in A} \mathbf{z}_i \phi_i p_i^{-2} \mathbf{z}_i' \right]^{-1} N^{-1} \sum_{i \in A} \mathbf{z}_i \phi_i p_i^{-2} e_i. \quad (13)$$

Under assumptions (design consistency standard assumptions)

$$N^{-1} \sum_{i \in A} \mathbf{z}_i \phi_i p_i^{-2} \mathbf{z}_i' = N^{-1} \sum_{i \in U} \mathbf{z}_i \phi_i p_i^{-1} \mathbf{z}_i' + O_p(n^{-1/2}).$$

Write

$$N^{-1} \sum_{i \in U} \mathbf{z}_i \phi_i p_i^{-1} \mathbf{z}_i' = \mathbf{M}_{zz,N}.$$

Use the same argument to expand the  $N^{-1} \sum_{i \in A} \mathbf{z}_i \phi_i p_i^{-2} e_i$  term. Then the expansion of (13) is

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_N + \mathbf{M}_{zz,N}^{-1} N^{-1} \sum_{i \in A} \mathbf{z}_i \phi_i p_i^{-2} e_i + O_p(n^{-1}).$$

For construction of confidence intervals for  $\bar{y}_N$  it is enough to consider the variance of the linearized term. Therefore consider in the notation of Särndal, Swensson, and Wretman (1992),

$$\text{AV}(\bar{y}_{\text{reg}}) = \bar{\mathbf{z}}_N' \mathbf{M}_{zz,N}^{-1} V(\bar{\mathbf{b}}_{\text{HT}} | F_N) \mathbf{M}_{zz,N}^{-1} \bar{\mathbf{z}}_N$$

where

$$\mathbf{b}_i = \mathbf{z}_i \phi_i p_i^{-1} e_i.$$

The variance of the HT estimator for the mean of  $b_i$  under Poisson sampling is

$$\sum_{i \in U} (1 - p_i) p_i^{-1} \mathbf{b}_i \mathbf{b}_i'.$$

Next apply that  $\phi = 1 - p_i$  to obtain the asymptotic variance approximation to the linearized part of  $\bar{y}_{\text{reg}}$

$$\text{AV}(\bar{y}_{\text{reg}}) = \bar{\mathbf{z}}_N' \mathbf{M}_{zz,N}^{-1} \sum_{i \in U} (1 - p_i)^3 p_i^{-3} \mathbf{z}_i e_i^2 \mathbf{z}_i' \mathbf{M}_{zz,N}^{-1} \bar{\mathbf{z}}_N.$$



The variance estimator is obtained by replacing the population totals with HT estimators under Poisson sampling and incorporating a degree of freedom correction to the front of  $n/(n-s)$  due to the small sample size.

## References

- Beaumont, J.-F., and Bocci, C. (2008). Another look at ridge calibration. *Metron*, 66, 1, 5-20.
- Chambers, R.L. (1996). Robust Case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.
- Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.
- Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.
- Fuller, W.A. (1981). An empirical Study of the ratio estimator and estimators of its variance: Comment. *Journal of the American Statistical Association*, 76, 78-80.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.
- Fuller, W.A. (2009a). Some design properties of a rejective sampling procedure. Forthcoming *Biometrika*.
- Fuller, W.A. (2009b). *Sampling Statistics*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Ikasi, C.T., and Fuller, W.A. (1982). Survey design under a regression population model. *Journal of the American Statistical Association*, 77, 89-96.
- Matei, A., and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21, 543-570.
- Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.
- Rao, J.N.K., and Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, D.C., 57-65.
- Rousseau, S., and Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Technical report, INSEE, Paris.
- Royall, R.M., and Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance – An empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- Royall, R.M., and Herson, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.
- Särndal, C.-E. (1980). On  $\pi$  inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag, Inc.
- Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer Science+ Business Media, Inc.
- Tillé, Y., and Matei, A. (2005). The R package Sampling. *The Comprehensive R Archive Network*, <http://cran.r-project.org/>, *Manual of the Contributed Packages*.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.



# The multidimensional integral business survey response model

Mojca Bavdaz<sup>1</sup>

## Abstract

Knowledge of the causes of measurement errors in business surveys is limited, even though such errors may compromise the accuracy of the micro data and economic indicators derived from them. This article, based on an empirical study with a focus from the business perspective, presents new research findings on the response process in business surveys. It proposes the Multidimensional Integral Business Survey Response (MIBSR) model as a tool for investigating the response process and explaining its outcomes, and as the foundation of any strategy dedicated to reducing and preventing measurement errors.

Key Words: Accuracy; Data collection; Economic statistics; Business survey; Measurement error.

## 1. Introduction

Measurement errors represent the gap between an ideal measurement and the obtained survey response (Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau 2004). To efficiently prevent or reduce the occurrence of measurement errors, it is necessary to know how the process of responding to survey questions evolves and what influences its course. Because work to reduce errors in business surveys has traditionally focused on sampling, frame, and nonresponse errors and, to a lesser extent, on measurement errors (Willimack, Lyberg, Martin, Japac and Whitridge 2004), knowledge of measurement errors and the underlying causal mechanisms is still largely limited in business surveys. This article attempts to fill that gap.

Most studies that examine the causes of measurement errors in business surveys are a product of pretesting research. As a result, most such studies are hypothetical (*e.g.*, Morrison, Stettler and Anderson 2002) or tentative (*e.g.*, Phipps, Butani and Chun 1995) as opposed to being based on actual data collection (*e.g.*, Hak, Willimack and Anderson 2003). The abundance of pretesting results, which are usually bound to a particular survey, contrasts with the scarcity of quality assessment research (*e.g.*, Giesen and Hak 2005) and with the shortage of generalization and linkages to the response process. Many studies focus on a particular aspect of the response process. For instance, Ponikowski and Meily (1989) examined the availability of data that business surveys require; Ramirez (1996) investigated respondent selection in business surveys; Jenkins and Dillman (1997) considered the design of business questionnaires; O'Brien (2000) and Willimack (2007) explored the respondent's role in the establishment survey response; Greenia, Lane and Willimack (2001) concentrated on business perceptions of confidentiality and on the closely connected issue of data sharing among statistical organizations; and Willimack (2003) exposed comprehension issues. Recently, more

attention has been dedicated to the development and testing of electronic business questionnaires (*e.g.*, Snijkers, Onat and Visschers 2007) and their editing (*e.g.*, Nichols, Murphy, Anderson, Willimack and Sigman 2005), while more frequent complaints about the costs that statistical reporting imposes on the business community have triggered research on the response burden (*e.g.*, Hedlin, Dale, Haraldsen and Jones 2005).

The first study to systematically address the entire response process in establishment surveys was a general model of the survey response process for factual information, which Edwards and Cantor (1991) presented. Biemer and Fecso (1995) combined the cognitive model of Edwards and Cantor's (1991) survey response with a statistical model that tried to quantify measurement errors by their sources. Another attempt to grasp the entire response process in business surveys was made in 1998-1999, when the U.S. Census Bureau conducted unstructured qualitative interviews on statistical reporting. The study served as a basis for two business survey response models: the hybrid response model for establishment surveys by Sudman, Willimack, Nichols and Mesenbourg (2000) and the complete model by Willimack and Nichols (2001). Most recently, Lorenc (2006) suggested examining the entire response process on the basis of the idea of socially distributed cognition and using an establishment as a unit of observation.

These models identify many essential aspects of the response process in business surveys and offer some concepts for them, but they treat many issues only partially. This was an incentive for a comprehensive study of the response process of a selected business survey making possible further development of the business survey response model. This article presents the Multidimensional Integral Business Survey Response (MIBSR) model and discusses its contributions.

1. Mojca Bavdaz, Faculty of Economics, University of Ljubljana, Slovenia. E-mail: mojca.bavdaz@ef.uni-lj.si.



## 2. Empirical study

The aim of the empirical study was to build a conceptual framework of the response process – a response model – by examining from start to finish the actual response process to a typical business survey in a real business environment. The qualitative research interview was the primary method of investigation. The method was implemented using various techniques (mainly retrospective probing and ethnographic interviewing but also thinking aloud), two modes (in person and by telephone), and different interviewees (people from the participating business, questionnaire administration experts from the statistical organization, and subject-matter experts). In some cases on-site observation and analyses of micro data complemented those techniques. Considering all the variables, a range of approaches had to be developed (for more details, see Bavdaž 2009). On-site visits were arranged around two consecutive deadlines for the questionnaire's completion in 2005. An attempt was made to contact all key people involved in the response process.

The selected survey – the Quarterly Survey on Trade – was a business survey conducted by the Statistical Office of the Republic of Slovenia on a sample of approximately 1,600 legal units performing trade activities. It had classic characteristics of business surveys: a recurring mandatory governmental mail survey. Its instrument was an eight-page paper questionnaire and instruction and classification booklets. The questionnaire consisted of an introductory text and four sections, one referring to the business as a whole and the other three each referring to one kind of trade activity (commission trade, wholesale, and retail). All sections asked for sales and employment data. In addition, there were questions on sales breakdowns, stock, activity codes, and size and number of stores. Nonresponding units received up to three reminders and, ultimately, a telephone call. The final response rates were generally high, greater than 90%. Major deviations and inconsistencies discovered during editing procedures also required telephone calls to businesses.

The final sample in this study consisted of 28 businesses required to complete the Quarterly Survey on Trade. Previous studies resulting in models of the response process applicable to business surveys were based on small samples as well: 24 establishments (Edwards and Cantor 1991), 30 large multiunit companies (Sudman *et al.* 2000; Willimack and Nichols 2001), and 7 schools (Lorenc 2006). This is consistent with exploratory interview studies, which tend to have small sample sizes of “around  $15 \pm 10$ ” (see Kvale 1996, page 102). The selection of businesses aimed to cover the heterogeneity of response processes. Because business size can be defined as the single most important business

characteristic that is assumed to influence or be related to the characteristics of the response process (*e.g.*, O'Brien 2000), businesses were selected from all size classes.

Several measures boosted the validity of the research design. The businesses were selected from different size classes, including some of the largest ones in trade but also some from nontrade primary business activity. A few businesses refused to cooperate, mainly because of the work overload. Nevertheless, caution is necessary when applying findings to nontrade and overworked businesses. The study included people with different roles in the response process. Substantial effort was made to obtain participation and organize visits during the time the respondents were completing the questionnaire or right afterward so as to minimize the loss of information from their memory. The short time lags that occurred in some cases did not seem to be so damaging for remembering a frequently repeated and well-documented process, given the advance announcement of the impending on-site visit. Interview questions directed respondents to report how they last filled out the questionnaire (*e.g.*, when the books closed that month, how much time they spent, who signed the form and how fast), and respondents generally supported their reports by data from paper and electronic documentation they used to fill out the questionnaire. All this helped distinguish their last engagement from the usual one.

The interview as the primary research method was in some cases combined with observation. The interviews were tape-recorded and transcribed. More repeating patterns emerged as the fieldwork progressed, though diminishing returns of each consecutive on-site visit were noted toward the end of the fieldwork. The findings from the on-site visits were compared with the observations of the survey staff and subject-matter experts, quantitative data (where available), and previously published research. Alternative explanations were considered. Last but not least, the selection of a typical business survey made the generalization to other business surveys more plausible. As Yin (2003) suggests, all steps in the research were carefully documented to establish a chain of evidence and ensure high reliability of findings.

## 3. The MIBSR model

### 3.1 Presentation of the model

One of the main study results is the Multidimensional Integral Business Survey Response (MIBSR) model, which integrates previous research findings and new findings from my empirical study. The MIBSR model explicitly distinguishes between processes occurring at the individual level and others taking place at the organizational level, which is the business level in this case (see Figure 1). The cognitive

processes of comprehension, retrieval, judgment, and response occurring at the individual level are taken from Tourangeau's (1984) response model. They reflect the mental processes of people involved in the survey response that relate to the actual answering of particular survey questions as compared to the processes that refer to the organization, information support, and authorization of such answering, which occur at the business level. Contrary to the typical situation of surveys of individuals, parts of the process, such as requesting data from another participant or retrieving data from business records, are visible through participants' physical actions. By using the survey level, the MIBSR model also allows for the possibility of conceptualizing the response process over several implementations of a survey or over several surveys (indicated by the arrows in Figure 1).

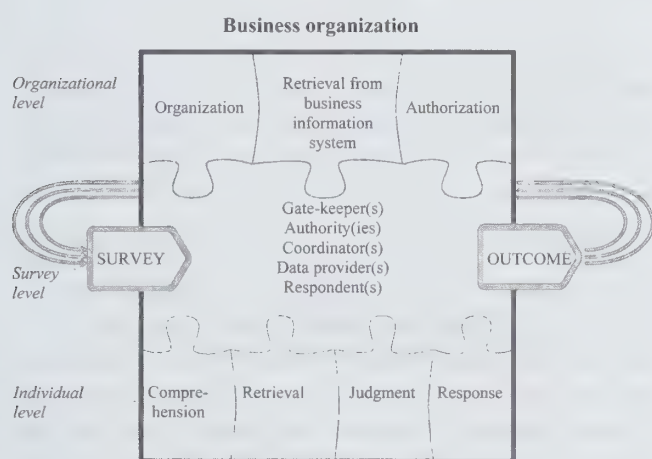


Figure 1 MIBSR model

The survey response task may involve several business participants who can enter and exit the response process at various points in time; but for the sake of clarity and simplicity, they are all depicted together. Business participants take part in organizational processes while going through their own cognitive processes; thus, they are a unifying link between processes at the individual and organizational levels. They may adopt one or more of the roles with a different influence on the response process, namely a gate-keeper (e.g., a receptionist, boundary-spanning unit), an authority, a response coordinator, a data provider, or a respondent. Although Figure 1 presents participants from a single business organization, successful completion of the task may require either the participation of people who provide outsourced activities or communication with survey staff.

The response process is triggered when the survey instrument crosses the business's boundaries. The MIBSR

model addresses the business response to a survey request presupposing a positive decision about participation in the survey. The examination of this decision, potentially leading to nonresponse, goes beyond the scope of this article even though it represents a natural introduction into the response process and may influence its course. The model suggests the most typical sequence of processes, although in practice some may be left out, repeated, or occurring in a different sequence. The following sections focus only on elaborated and newly added insights into the response process.

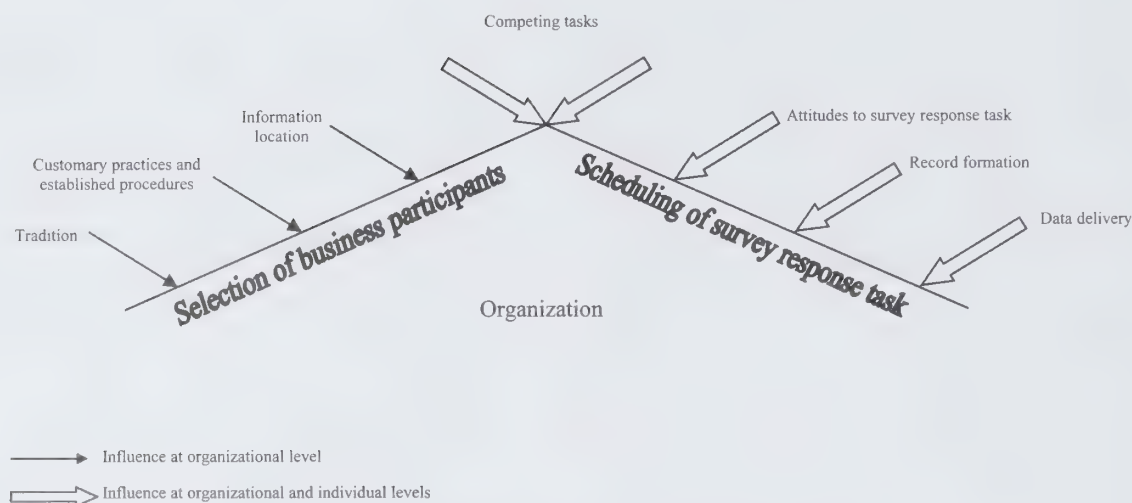
## 3.2 Organizational level

### 3.2.1 Organization of the survey response

Participation in a survey generally entails some preparatory activities due to work distribution and specialization in organizations. It requires an answer about who will perform the survey response task and when it will be done; both answers provide clues about how the task will be carried out. The study provided evidence that the two steps could be intrinsically linked. In fact, the selection of people for the survey response may itself indicate the priority assigned to the task in the organization. For instance, in some accounting firms and larger businesses, chiefs performed the task themselves, although they could have delegated the task, which may indicate a certain importance of the task, while the fact that many respondents received the task as novices may indicate its low priority. In contrast, priorities at the individual level were not always consistent with priorities at the organizational level. For instance, even if tax reporting gained higher priority than statistical reporting at the organizational level, this was irrelevant for a survey respondent not involved in tax reporting. I therefore examined the selection of business participants and the scheduling of the survey response task together within the organization of the survey response. The result is an expanded list of factors potentially influencing the organization of the survey response task (see Figure 2).

Tradition, customary practices, established procedures, and information location mainly influence the selection of business participants, which is an organizational matter, while other factors operate at both the organizational and the individual levels. Tradition dictates reliance on previous participants in recurring surveys when the same people repeatedly participate in the response process of the same (longitudinal) survey. Some study respondents claimed they had been "filling it out for years." Some had been filling it out since they started the job or since a colleague retired, went on a longer sick leave, left the job, and so on.





**Figure 2** Factors influencing the organization of survey response

Many processes in organizations draw on customary practices and established procedures, which leads to the selection of the usual participants. This means that even when a new survey request reaches the business, the business will likely proceed in the same way as with previous survey requests because of the relatively stable distribution of work. In fact, some of the respondents in this study explained that the survey questionnaire would often be directed to the same department or person, who usually replied to such requests even if no formal policy on surveys existed. As one respondent clarified, “They prefer to bring them to me – this is the only policy.” Some respondents knew which types of surveys they received, saying, for instance, “I’m doing all statistics except wages,” or “I’m doing all statistics, also for the Bank of Slovenia, except Intrastat.” Even in larger businesses, the same person often filled out several different survey questionnaires; one person completed all survey questionnaires that required financial data, be it for the Bank of Slovenia, the Statistical Office, or the Agency for Public Legal Records; others provided a list of specific surveys that they would complete, such as surveys on investments, fixed assets, value added, and so on.

Information location is an essential factor that influences the selection of business participants from the perspective of measurement errors. It refers to sufficient knowledge to provide an accurate survey response, including adequate access to records, if necessary. In this study, many respondents expressed that they had been chosen because of their access to data, for instance, “I have the data and I know how to retrieve them.”

Competing tasks relates to the assignment of people and order to the tasks. It usually influences the choice of business participants at the organizational level when

alternative possible participants are compared, as well as the scheduling of the survey response task at the individual level when the priorities of a participant’s several tasks are considered. Study respondents in several, mainly smaller businesses agreed that they give low priority to the survey response task when they schedule their work: “VAT (value-added tax), debt recovery, bookkeeping . . . all has priority over statistics.” Another respondent said that she “wouldn’t think of doing the survey on the day all the book entries are done” but instead checks “the balance sheet, . . . liabilities, how the payments stand, how much debt there is, the financial situation.” Another explained the work process as “internal reporting first, current affairs next, statistical reporting afterwards.” In a few larger businesses, however, respondents said that they completed survey questionnaires as soon as data became available or final.

Similarly, attitudes to the survey response task can be examined at the organizational level through formal policies on surveys and the informal reactions of authorities as well as individual perceptions. Businesses in this study did not have any formal policies on surveys, though the discourse of authorities in some companies indicated their negative attitudes: “it’s only statistics; prepare something.” Organizational attitudes may affect the organization of the survey response, through potential consequences for the business, particularly opportunity costs, penalties, and damage to the public image. Most participants expressed a negative attitude toward surveys, describing them as “a necessary evil” and “redundant” or “additional” work. Individual attitudes toward surveys may contribute to the early, timely, or late scheduling of the task; they may also influence an individual’s inclusion or exclusion in the survey response task.



Record formation and data delivery are primary in the scheduling of the response tasks. The timing of record formation determines when the records with required data about the business were created and took on the acceptable or desirable form, especially when the data become final. Respondents in larger businesses and businesses with foreign ownership typically referred to internal deadlines for “closing the books” or the VAT submission deadline. Data delivery is relevant in those cases where the participant must rely on other people to deliver required data. This particularly applied to accounting firms in this study. However, the timing of record formation and data delivery may vary by the kind of data requested, so that the latest record formation and the latest data delivery, eventually, determine the actual scheduling. For instance, some respondents explained that more time was necessary to get the correct value of stock because of lags in recording incoming invoices as compared to sales figures.

After the organization of the survey response task, the task can be realized, though it is sometimes necessary to further refine the selection of business participants or the scheduling to provide for all requested items, absence from work, and other circumstances.

3.2.2 Retrieval of information from the business information system

The capacity of the business information system (BIS) is the key factor that influences the response process and its outcome in business surveys. The BIS does not consist of the technological element only; it also includes people (Avison and Elliot 2006). The human capacity of the BIS relevant for the business survey response is mainly reflected in cognitive processes at the individual level (see section 3.3), while its technological capacity is determined through business records at the organizational level. The study showed that formation of business records depends on internal and external factors, though the line between the two groups is blurred (see Figure 3).

External factors – legal obligations, standards, and benchmark practices – are imposed on companies from the environment and dictate the content of business records through cogency or the threat of sanctions. Legislation, regulations, and other forms of power with the law set out legal obligations. With respect to that, study respondents mainly mentioned mandatory compliance with accounting standards and the requirements of tax authorities. The latter could refer to the business as a whole (*e.g.*, VAT reports) or to particular items (*e.g.*, excise duties on tobacco products). Other mandatory requirements may relate to contributions, securities, insurance, environmental issues, and so on. Participants usually noted the compulsory character of governmental business surveys, although the lack of

sanctions for nonresponse or a late response made some participants question this; furthermore, changing record formation for statistical purposes only was unthinkable to most study participants. Standards are a softer form of external factors: they are not mandatory, but are expected to be followed in most cases. Two examples from the study include the use of a classification based on the European Article Number barcode standard and recommendations from accounting authorities. The study suggested that standards were not used in the case of specific reasons; for instance, the information systems of the smallest retailers did not support barcode use. Benchmark practices are the least influential group of external factors. They refer to good examples of practice that have gained some recognition and authority by reputation (and not by law or institutional power). For instance, some study respondents mentioned obsolete software versus current standards, while others stressed powerful capabilities of their software and its positive influence on data provision.

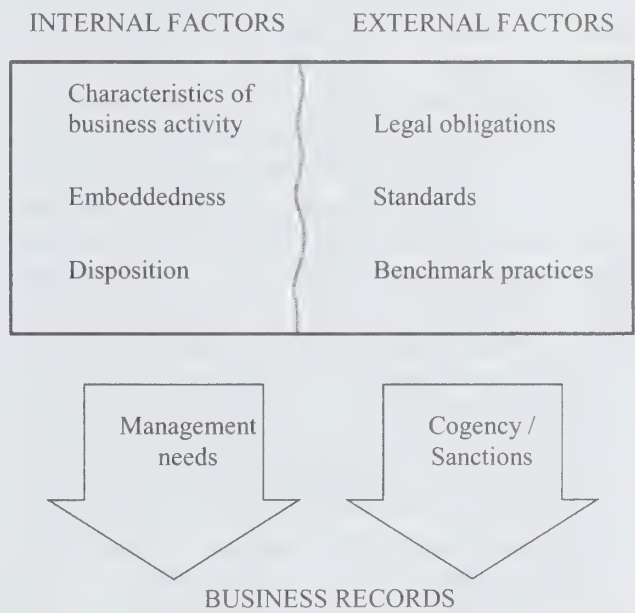


Figure 3 Factors of record formation

External factors drive data homogeneity and comparability in business records across companies, at least within similar economic activities. They provide the framework in which companies develop their own solutions for business records according to internal factors unless adhering to compulsory requirements more than fully satisfies data needs for running the business as was the case in small, local companies. Internal factors of record formation include characteristics of business activity, such as the size, type, and diversity of the business activity;

embeddedness in the business environment; and the disposition to forming records.

The size of the business activity plays a crucial role in record formation because it leads to a differential overview of an activity. In the study, most larger companies had an abundance of data. Business records provide information that cannot be gained from participation or observation only. That said, the size of the business activity is relative, especially if the size is observed only within legal boundaries or national borders. Therefore, it is better to speak about the embeddedness in networks of various kinds. In the study, for instance, a couple of smaller businesses had a foreign owner that demanded comprehensive reports to overcome the distance and manage the business remotely, and another small business had to use the sophisticated software of a business partner because it was its major supplier. The study also showed how different types of activities influenced the kind of available records; for example, wholesale businesses that typically put recipients on their invoices had more information on their buyers than businesses in retail that typically issued receipts without indicating the name. High diversity of business activities also is a major challenge for record formation in most businesses; in general, smaller businesses had renounced the use of detailed records and were forced to make estimates instead. Last, disposition refers to the prevailing attitudes of people in the business to various aspects of record formation, such as the inclination toward data, information technology, and change. Some businesses relied heavily on evidence-based decision making and thought highly of data; others showed enthusiasm for the possibilities of information technology, but a few others saw no usefulness in data.

Factors of record formation influence the availability of data in business records and their compliance with survey definitions. Data availability appears at the intersection of technological and human capacity in the business; knowledge is required to extract data from the BIS conditional on their existence. Several levels of answer availability in the BIS apply to survey questions (see Figure 4); their naming was inspired by the determination of cognitive states in Beatty and Herrmann (2002) and is in principle consistent with that proposed by Lorenc (2007):

- (a) A datum is accessible – the required answer may be readily available. In this study, a typical example is total sales revenue, which is readily available to a person in accounting, or the number of employees, which is readily available to a person in the personnel department.
- (b) A datum is generable – the required answer is not readily available to any person; the available data represent a basis for generating the required answer

through manipulation. In the study, for instance, sales revenue in a particular trade activity was not always readily available, but it was possible to derive the exact figure by consulting two separate records (e.g., the general ledger and commercial records).

- (c) A datum is estimable – the required answer is not readily available to any person; the available data represent an approximation of the required answer or a basis for estimating the required answer through manipulation. In the study, a sales breakdown by commodity groups (e.g., food, beverages, clothes, footwear) was often estimated by recategorizing available groups; however, those categories were sometimes too aggregated or too diverse to allow for an exact match (e.g., Christmas products, Easter gifts, discontinued products).
- (d) A datum is inconceivable – no available data lead to the required answer or its approximation; some bases for generating or estimating the required answer exist but require an unimaginable effort to produce it. For instance, a company would have to classify more than ten thousand invoices monthly to arrive at an exact breakdown of sales by kind of buyers.
- (e) A datum is nonexistent – there are no bases for estimating the required answer. In the study, a cash-and-carry store could not distinguish between different kinds of buyers because they issued the same kind of nameless invoices to all customers, companies and individuals.

Because data availability varies across people in a business, it may be useful to determine answer availability at the individual level. In this case, a distinction has to be made between an answer that someone can obtain directly and an answer that they can access only through another person.

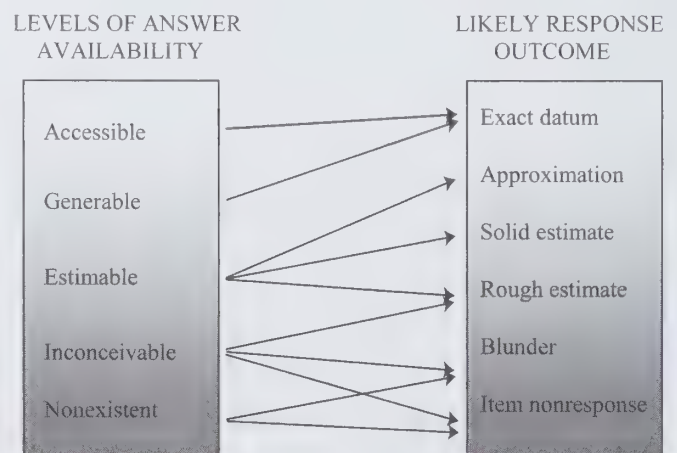


Figure 4 Levels of answer availability and likely response outcome



The final response outcome is conditional on the level of answer availability and may range from an exact datum to item nonresponse (see Figure 4). A measurement error occurs whenever the response outcome deviates from the exact datum. When a datum is accessible or generable, the response outcome is likely to be an exact datum, although the possibilities of committing a measurement error increase if data have to be accessed through other people or manipulated. When a datum is estimable, the response outcome may be an approximation with a negligible measurement error or an estimate with a minor or substantial measurement error. An inconceivable datum may, at best, lead to a rough estimate. When respondents have no adequate bases to provide a response, they may make wild guesses resulting in blunders or skip the question, which leads to item nonresponse.

### 3.2.3 Authorization of the business response

Authorization is the final opportunity for corrective actions before the business response is forwarded to the survey organization and documentation archived. Most businesses in this study found this organizational step inconsequential and even skipped it. In more than half of businesses, respondents signed the questionnaire themselves because “they have the mandate to sign such things” and “the director is very rarely present” or “does not deal with such things.” Still, even in those cases, some respondents mentioned that the director had been informed about that procedure. In several businesses, the superior signed the questionnaire for the sake of formality and no verification procedures were in place because “the director trusts us” or “doesn’t have the necessary data,” or because “we work this way.”

A superior was typically present in the largest companies, through formal authorization or informal notification. Internal verification was rare, which could be the consequence of preceding consultations with the superior. Accounting firms usually delivered the completed questionnaire to the business for signature, though businesses sometimes also signed the blank questionnaire in advance.

### 3.3 Individual level

Given the level of answer availability in the BIS, it rests on the performance of cognitive processes and accompanying physical actions (especially interaction with computers) at the individual level to determine the final response outcome. The MIBSR model proposes that three inherently linked types of knowledge are relevant for these processes: knowledge of business reality, knowledge of record formation, and knowledge of business records (see Figure 5). Although it may be difficult to disentangle the three types of knowledge in practice, the study seems to suggest that every type is particularly influential for one kind of cognitive process.

The division of cognitive processes into comprehension, retrieval, judgment, and response derives from Tourangeau’s (1984) response model. In business surveys, these processes may not be defined as easily as in surveys of individuals because the initial organization of the response may involve only a brief and superficial consideration of the survey task with barely any impact on the later response process or a thoughtful reflection on the questions. The study mainly focused on respondents’ cognitive processes because it is their task to answer survey questions. Nevertheless, observations of other business participants are provided where available.

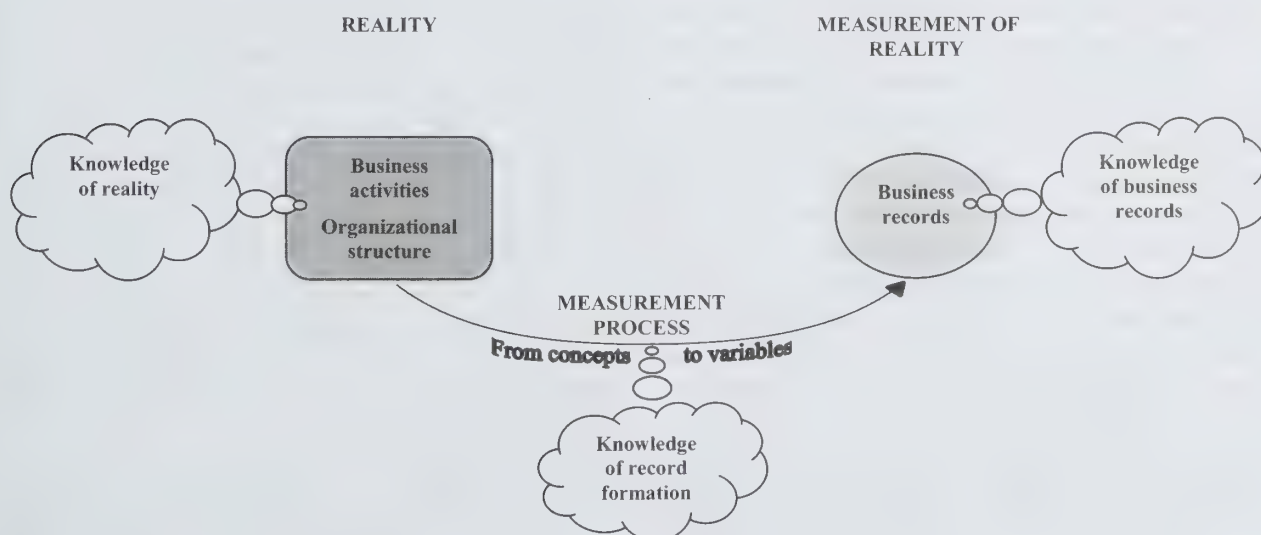


Figure 5 Knowledge relevant to the business survey response



### 3.3.1 Comprehension

In comprehension processes, respondents interpret the survey request for data, which usually is in the form of labels instead of questions. The MIBSR model suggests that, for comprehension processes, knowledge of business reality is particularly important. Business reality refers to the activities the business performs to subsist and to the division of work across locations and individuals. Knowledge of business reality thus presupposes acquaintance with every aspect of the business: who does what, what activities the business is involved in and how they are carried out, how decisions are made, why the business situation is as it is, how it evolved through time, and so on. Because larger businesses tend to be complex with technical and social divisions of labor, establishment of branches, organizational hierarchy, and decision-making structure (Tomaskovic-Devey, Leiter and Thompson 1994), it can be expected that fragmentation of the knowledge of the business's reality increases with business size.

This knowledge is essential in establishing whether survey questions are applicable to the business and providing correct answers afterward. In fact, no business in the study filled out all survey items. Respondents had to fill out only sections that applied to the kinds of trade they performed. Survey questions also required them to select applicable commodity groups, kinds of employment, kinds of buyers in wholesale, kinds of payment in retail, and so on. The required knowledge of business reality was occasionally specific: one respondent, for instance, needed information about the relationship between the company as the franchisor and their franchisees to avoid double counting or skipping some items across the businesses.

A major obstacle to using knowledge of business reality for correctly understanding survey questions was the incomprehension of economic and accounting concepts or their confounding with other concepts. For instance, one respondent had problems distinguishing between the concept of trade, which includes repackaging of goods, and the concept of production, which entails some transformation of goods beyond repackaging; a few respondents pondered over trade rendered on a commission basis because their activity was trade but accounting treated it as a service; many respondents associated retail with a store rather than with individuals as final consumers, regardless of the kind of buyer; one respondent defined wholesale as "everything that is not paid with cash" instead of linking it to nonfinal consumption; some respondents did not understand that "nontrade and nonmanufacturing organizations" were service providers; others did not understand the difference between merchandise and material, because the latter is an input to production (not trade) in accounting terminology

and takes on another meaning in colloquial language, such as construction or building material.

Study respondents often used their own definitions to interpret survey questions. The same is true for those business participants who provided data on request without actually seeing the questionnaire and/or instruction booklet. This, for instance, happened in a few larger businesses where data providers completely relied on their own definitions of the sales space when providing data on store distribution by size of the sales space because additional explanations were given only in the instruction booklet.

### 3.3.2 Retrieval

In retrieval processes, the data and information required for the survey response are located and brought forth. In business surveys, the data usually reside in business records, not in memories, but knowledge is crucial for their extraction and interpretation. The retrieval thus mainly rests on knowledge of the business records, which refers to the contents and location of business records in the business and the possibilities of data access, including familiarity with applications and the people in charge of them.

Study respondents mainly exhibited good knowledge of the business records they worked with. In a couple of businesses where superiors participated in the response process, the superiors were not abreast of all details of the records and had an assistant perform the retrieval—but they had excellent insight into the business reality and knew how it converted into records. Even perfect knowledge of the business records, however, did not always suffice for exact answers. When the business records did not register all necessary data, knowledge of the business reality became critical for making correct inferences and good estimates. This sometimes happened in larger businesses and accounting firms where respondents knew the records very well, including the chart of accounts and its codes, but knew the assortment of merchandise only vaguely. As a result, they had to use estimates when classifying sales by commodity groups, as their acquaintance with the business activity was incomparable to a comprehensive, firsthand insight of sales personnel. In smaller businesses, lack of necessary data in records sometimes meant complete reliance on memory instead of records; a respondent, for instance, arrived at employment in wholesale by retrieving the number of people in relevant workplaces, namely chauffeurs, people who worked in the warehouse, salespeople, and office clerks.

### 3.3.3 Judgment

Judgment refers to the compilation of all retrieved data and information to formulate an answer. In this study, it frequently entailed some data manipulation or handling,

such as summation, balance with a residual, recategorization, and application of proportions. Judgment is mainly supported by knowledge of record formation. This knowledge provides information on how the business reality translates into business records and ensures that captured data are not considered isolated figures, codes, or words but take on a certain meaning representing the processes and objects measured. It therefore represents a link between knowledge of the business reality and knowledge of the business records (see Figure 5). Its importance was, for instance, noted during the observation of a respondent who was filling out the questionnaire and had to struggle with an inconsistency in the retrieved sales data. To identify the mistake, she systematically analyzed nonsales activities in the observed period and the correctness of their encoding in the records to finally discover a transaction that should not have been included in the sales figures.

However, lack of knowledge could not explain some judgments with an unfavorable response outcome, so the study looked more closely into principles that guided judgment. Among the most pervasive principles encountered in the survey response process under study was the principle of continuity, which advocates the use of the same response strategy in recurring surveys – even if this leads to errors. Continuity was sometimes considered within a year but also across years. It seemed to be strengthened by the lack of negative feedback from the statistical organization and its presumed satisfaction with the data. The study identified several respondents who used detailed procedures of calculation that were quite obsolete. A respondent even erroneously left out the section of commission trade but would not change the procedure during the year to avoid disrupting the reported data.

Two other principles were identified in relation to the principle of continuity: the principle of consistency and the principle of disregarding the exceptional. The principle of consistency implies use of the same or similar response strategies in the same survey questionnaire. For instance, a respondent who attributed various items of merchandise to only one commodity group in wholesale did the same in retail; a respondent who estimated wholesale turnover from VAT figures used the same approach to retail turnover, and so on. The principle of disregarding the exceptional implies ignoring new, one-off, or temporary activities. For instance, a study respondent inadvertently reported a temporary activity not reported in the questionnaire; another confessed the exclusion of new activities from reporting because their success was uncertain. The question, however, is how to set boundaries on the novelty and on the temporariness and when precisely such activities become representative of the business.

The principle of disregarding the exceptional is also related to the principle of disregarding the marginal, which advises ignoring those activities that are perceived as marginal to the business. For instance, some study respondents disregarded some items in sales breakdowns if they represented less than one percent of activity. The impact of the principle depends on the use of the collected data. It should be inconsequential if the aim is to estimate national totals or change. However, sales of a specific commodity group may be marginal to a large business but not marginal for the market of that commodity group.

The business perspective principle advocates the priority of the business perspective as compared to a statistical request. In the study, data on existing organizational units were judged acceptable despite their divergence from the required units; data on various packages (*e.g.*, a newspaper supplemented with a book) that were relevant from the business perspective were not disentangled for statistical purposes.

### 3.3.4 Response

The response component refers to the processes of mapping a judgment onto a response category and editing the response (Tourangeau, Rips and Rasinski 2000). In business surveys, mapping usually translates into matching available data from the BIS with response categories offered, which provides room for a specific form of measurement error: misclassification. For instance, when respondents had problems fitting available sales data into the provided classification scheme, they often chose the closest category, the main category, or the category “other.”

The study also identified the presence of editing processes that show different aspects of business sensitivity. Some study respondents checked whether their selection of the decisive activity code was consistent with their registered activity, which may show a fear of nonconformity with administrative requirements. Not reporting people who helped in family businesses may reveal tax evasion. Although many respondents agreed that the data they reported in the questionnaire were considered confidential, there was scarce evidence of hindrance for disclosing the data to the statistical organization (*e.g.*, not reporting detailed data on newly introduced activities).

### 3.4 Survey level

The MIBSR model introduces the possibility of conceptualizing the response process over several implementations of a survey or over several surveys. It thus conceptually enables the observation of how the elements of survey design, which is under the control of survey organization, influence the response process.



The study focused on the impact of recurrence on the response process. In repeated administrations of the survey to the same business, the organization of the survey response became less relevant or irrelevant if it was a perfect replica of the preceding administration. The cognitive processes at the individual level were characterized by routine when the same business participants performed them. Many respondents admitted that they had not read the whole questionnaire, let alone the instructions in a repeat questionnaire. This also occurred in businesses that agreed to be observed while completing the questionnaire: after respondents gave the questionnaire a swift scan for any changes, they plunged into the retrieval processes based on the previously completed questionnaire or on other documentation and supporting notes. The comprehension step was thus performed superficially and pertained more to understanding completion of the previous questionnaire than it did to understanding survey requests. The retrieval procedures followed the previously established course and exhibited learning-curve effects. The respondent's judgment clung to the initial approach and was unlikely to change. The recurrence frequently loosened up a respondent's supervision and reduced the importance of the authorization or even omitted it.

Given the appointment to the survey task of the same people or usual units in the business, many of them sooner or later had contact with survey staff, despite the common self-administrative mode of data collection in business surveys. Such contact could occur early in the response process and influence the respondent's comprehension and judgment. This was rarely the case in the study; only a few respondents asked for explanations the first time they participated in the survey and another respondent asked for help when the business's activity changed. Contacts in which respondents requested postponement of the deadline did not seem to influence the subsequent response process, though the same could not be claimed for respondents who resisted participation. All other contacts happened during a follow-up when the response process, or parts thereof, had to be performed again, which could result in an adjusted survey response. Although respondents mainly acknowledged the politeness of the survey staff, their calls signaled that something was wrong: a missed deadline, an item missing in the questionnaire, an inconsistency in the reported data. The rareness of such contacts made a significant impression on respondents because these contacts were often the only type of feedback from the statistical organization.

In contrast, respondents did not always appreciate a lack of feedback. They expected feedback from the statistical organization after they first participated in the survey, but this generally did not happen. The lack of reaction made

them confident in their approach, thus reinforcing the principle of continuity in their judgment. However, many respondents reported at least one piece of data that was not completely accurate (or not as accurate as they would expect the data should be) and they perceived the lack of complaints as satisfaction with bad data. Some respondents were convinced that the statistical organization knew about their business activity, which is why they rarely provided textual descriptions of seasonal oscillations. Given these observations, it is not surprising that several respondents expressed doubts about the accuracy of statistical data or questioned the accuracy of data that others provided. The right feedback may not only be important for that particular survey but also for participation in other surveys because it contributes to general perceptions on surveys and statistics.

#### 4. Discussion of model's contributions

The dominance of written communication between the survey organization and businesses has moved business participants away from the center of statistical production and reduced the possibilities of insights into the process of responding to survey requests and the causes of measurement errors. By studying the response mechanisms and influencing factors, response models help bring these insights out and design approaches that turn this knowledge into an advantage. This section discusses the contributions of the MIBSR model with respect to previous response models applicable to business surveys.

##### 4.1 Model construction

Two approaches were encountered in construction of previous models: adding some organizational steps to the core cognitive processes from Tourangeau's cognitive model of survey response (Biemer and Fecso 1995; Edwards and Cantor 1991; Sudman *et al.* 2000; Willimack and Nichols 2001) or using the organization as the unit of observation (Lorenc 2006). The MIBSR model explicitly links the processes to the level at which they occur: cognitive processes to the individual level and organizational processes to the organizational (in our case, the business) level. It also foresees the observation of the response process over several implementations of the same survey or over several surveys with different designs, which is particularly interesting for governmental surveys. By analyzing complex response processes at the appropriate level of observation, the MIBSR model sets up a framework that can also be used for quantitative modeling and experimental design.



## 4.2 Insights at the organizational level

Previous models treated initial organizational arrangements in the context of respondent selection (Biemer and Fecso 1995; Edwards and Cantor 1991) or in separate steps of respondent selection and the assessment of priorities, the latter ranking statistical reporting to the government lower than most other business reporting activities (Sudman *et al.* 2000; Willimack and Nichols 2001). They also identified several factors that influence respondent selection, especially the functional role, authority level, and position with regard to the information system (Edwards and Cantor 1991), knowledge of the information system, terms and definitions (Biemer and Fecso 1995), competing job responsibilities and access to the data (Sudman *et al.* 2000). The MIBSR model integrates all preparatory activities in the organization of survey response and suggests an expanded list of influencing factors. The organization of survey response now acknowledges that delegation of the task may also include selection of other business participants beyond respondents and that priority of competing tasks is just one of the factors influencing the task's scheduling.

All previous models have paid considerable attention to record formation. The MIBSR model suggests a different systematization and extension of factors of record formation, initially grouped into management, regulation, and standards by Willimack and Nichols (2001). Because it is generally unlikely that the requirements of statistical reporting are an actual factor of record formation, the MIBSR model may assist the survey organization in its endeavors to exert influence on record formation and eventually obtain requested data. Taking into account technological and human capacity of the BIS, the MIBSR model defines several levels of answer availability based on the extent to which the answer conforms to required survey definitions and proposes the likely response outcome. In authorization of the business response, the MIBSR model reiterates the possibility of internal verification that Sudman *et al.* (2000) and Willimack and Nichols (2001) propose for the release step. Authorization is more likely sought out when the survey response involves legally separate units and more formalized and centralized organizations.

## 4.3 Insights at the individual level

At the individual level, which deals with comprehension, retrieval, judgment, and response (Tourangeau 1984), the MIBSR model further elaborates on the knowledge relevant to cognitive processes. Willimack and Nichols (2001) emphasized personal knowledge for answers directly from memory and knowledge of the records. The MIBSR model suggests that a thorough understanding of the data in business records and their appropriate use in the survey response require knowledge of the whole chain of data

generation, from knowledge of business reality to knowledge of record formation and knowledge of business records.

As far as comprehension processes are concerned, Edwards and Cantor (1991) have acknowledged the problematic use of jargon, and Sudman *et al.* (2000) have pointed to the problematic deviation of required economic concepts from accounting standards. The MIBSR model goes even further to explain that the errors may result from a broader issue of incomprehension of economic and accounting concepts or their confounding with other concepts.

The MIBSR model identifies several principles that help understand the underlying judgment processes in business surveys, which are consistent with examples manifesting the principles of continuity and consistency by Sudman, *et al.* (2000) and Willimack, Nichols and Sudman (2002), respectively. These principles may also reflect satisficing (Simon 1957) or inertia. The use of inappropriate principles, especially the principle of continuity, is particularly strengthened by the lack of survey feedback.

In the cognitive processes of responding, the MIBSR model exposes the problem of matching in business surveys, thus adding to the rounding error that Sudman *et al.* (2000) discuss. It also integrates different aspects of business sensitivity that Edwards and Cantor (1991) have discussed as part of the communication step, and Sudman *et al.* (2000) have discussed as part of the release step. The model treats them at the individual level where the editing occurs if the data are indeed sensitive.

## 4.4 Insights at the survey level

Previous models have concentrated on a single occurrence of the response process in a particular business survey, while the MIBSR model extends to several occurrences and several surveys. Among the many dimensions at the survey level, the study systematically analyzed the impact of recurrence and contact with the survey staff on the response process, which represents a further elaboration of specific instances already mentioned in previous models in the context of retrieval, such as rehearsal of the look-up (Edwards and Cantor 1991) or documentation of previous completions supporting retrieval (Sudman *et al.* 2000). In addition, the MIBSR model allows for the presence of a contagious effect transmitting the experience in one business survey to other business surveys.

## 5. Conclusion

Survey organizations usually have to set aside a considerable amount of resources for processing survey data because the processes of responding to survey questions in the businesses are not performed satisfactorily. The MIBSR model provides further evidence on how the processes are

carried out and what influences them. It offers insights into the business perspective, which are valuable for efficiently seeking solutions to improve the processes and, consequently, reduce or eliminate measurement errors. The model may also serve as a framework for the documentation and systematization of existing and future knowledge on the causes of measurement errors in business surveys. It may be used as a preceding step of empirical studies on measurement errors and for a consistent explanation of empirical findings. Future research should continue with the application of the qualitative research methods to the study of particular dimensions of the response process, other business participants besides respondents and other kinds of business surveys. It should also embark on quantitative modeling of the response process and verifying the effectiveness of suggested improvements with experiments. Last, it should look into the interactions with other kinds of nonsampling errors.

### Acknowledgements

This article is an outcome of doctoral research. The author thanks the Statistical Office of the Republic of Slovenia for its co-operation and Learegar (University of Ljubljana), Lars Lyberg (Statistics Sweden, Stockholm University) and Jaak Billiet (Catholic University of Leuven) for their guidance and support. I also thank the associate editor and anonymous referees for their helpful comments on an earlier version of this article.

### References

- Avison, D., and Elliot, S. (2006). Scoping the discipline of information systems. *Information Systems: The State of the Field*, (Eds., J.L. King and K. Lyytinen). Hoboken: John Wiley & Sons, Inc., 3-18.
- Bavdaž, M. (2009). Conducting research on the response process in business surveys. *Statistical Journal of the IAOS*, 26, 1-14.
- Beatty, P., and Herrmann, D. (2002). To answer or not to answer: Decisions processes related to survey item nonresponse. *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little). New York: John Wiley & Sons, Inc., 71-85.
- Biemer, P.P., and Fecso, R.S. (1995). Evaluating and controlling measurement error in business surveys. *Business Survey Methods*, (Eds., B.G. Cox et al.). New York: Wiley-Interscience, 257-281.
- Edwards, W.S., and Cantor, D. (1991). Toward a response model in establishment surveys. *Measurement Errors in Surveys*, (Eds., P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman). New York: Wiley-Interscience, 211-233.
- Giesen, D., and Hak, T. (2005). The response process model in business surveys: Lessons learned by using a multi-method approach. *FCSM Conference Papers*, Federal Committee on Statistical Methodology.
- Greenia, N., Lane, J. and Willimack, D. (2001). Perceptions of confidentiality protection at statistical agencies: Some evidence from data on businesses and households. *Statistical Journal of the United Nations ECE*, 18, 309-314.
- Groves, R.M., Fowler, F.J., JR., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. Hoboken: Wiley-Interscience.
- Hak, T., Willimack, D.K. and Anderson, A.E. (2003). Response process and burden in establishment surveys. *Proceedings of the Section on Government Statistics*, American Statistical Association, 1724-1730.
- Hedlin, D., Dale, T., Haraldsen, G. and Jones, J. (2005). *Developing Methods for Assessing Perceived Response Burden*. Eurostat.
- Jenkins, C.R., and Dillman, D.A. (1997). Towards a theory of self-administered questionnaire design. *Survey Measurement and Process Quality*, (Eds., L.E. Lyberg et al.). New York: Wiley-Interscience, 165-196.
- Kvale, S. (1996). *InterViews: An Introduction to Qualitative Research Interviewing*. Thousand Oaks: Sage Publications.
- Lorenc, B. (2006). Two topics in survey methodology: Modelling the response process in establishment surveys; inference from nonprobability samples using the double samples setup. Doctoral dissertation, Department of Statistics, Stockholm University.
- Lorenc, B. (2007). Using the theory of socially distributed cognition to study the establishment survey response process. *Proceedings of the Third International Conference on Establishment Surveys, Montreal, Canada*, American Statistical Association, 881-891.
- Morrison, R.L., Stettler, K. and Anderson, A.E. (2002). Using vignettes in cognitive research on establishment surveys. *International Conference on Questionnaire Development, Evaluation and Testing Methods, Charleston*, American Statistical Association.
- Nichols, E.M., Murphy, E.D., Anderson, A.E., Willimack, D.K. and Sigman, R.S. (2005). Designing interactive edits for U.S. Electronic Economic Surveys and Censuses: Issues and guidelines. *Research Report Series (Survey Methodology 2005-03)*, U.S. Census Bureau.
- O'Brien, E.M. (2000). Respondent role as a factor in establishment survey response. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 1462-1467.
- Phipps, P.A., Butani, S.J. and Chun, Y.I. (1995). Research on establishment-survey questionnaire design. *Journal of Business & Economic Statistics*, 13, 337-346.
- Ponikowski, C.H., and Meily, S.A. (1989). Controlling response error in an establishment survey. *Proceedings of the Surveys Research Methods Section*, American Statistical Association, 258-263.
- Ramirez, C. (1996). Respondent selection in mail surveys of establishments: Personalization and organizational roles. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 974-979.
- Simon, H. (1957). *Models of man: Social and rational*. New York: John Wiley & Sons, Inc.
- Snijders, G., Onat, E. and Visschers, R. (2007). The annual structural business survey: Developing and testing an electronic form. *Proceedings of the Third International Conference on Establishment Surveys, Montreal, Canada*, American Statistical Association, 317-326.

- Sudman, S., Willimack, D.K., Nichols, E. and Mesenbourg, T.L. (2000). Exploratory research at the U.S. Census Bureau on the survey response process in large companies. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 327-337.
- Tomaskovic-Devey, D., Leiter, J. and Thompson, S. (1994). Organizational survey nonresponse. *Administrative Science Quarterly*, 39, 439-457.
- Tourangeau, R. (1984). Cognitive science and survey methods. *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, (Eds., T.B. Jabine, M.L. Straf, J.M. Tanur and R. Tourangeau). Washington, D.C.: National Academy Press, 73-100.
- Tourangeau, R., Rips, L.J. and Rasinski, K.A. (2000). *The Psychology of Survey Response*. Cambridge, England: Cambridge University Press.
- Willimack, D.K. (2003). Business respondents' perspectives on alternative employment arrangements and implications for employment statistics. *Proceedings of the Section on Government Statistics*, American Statistical Association, 4559-4570.
- Willimack, D.K. (2007). Considering the establishment survey response process in the context of the administrative sciences. *Proceedings of the Third International Conference on Establishment Surveys, Montreal, Canada*, American Statistical Association, 892-903.
- Willimack, D.K., Lyberg, L.E., Martin, J., Japac, L. and Whitridge, P. (2004). Evolution and adaptation of questionnaire development, evaluation, and testing methods for establishment surveys. *Methods for Testing and Evaluating Survey Questionnaires*, (Eds., S. Presser et al.). Hoboken: Wiley-Interscience, 385-407.
- Willimack, D.K., and Nichols, E. (2001). Building an alternative response process model for business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Willimack, D.K., Nichols, E. and Sudman, S. (2002). Understanding unit and item nonresponse in business surveys. *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little). New York: John Wiley & Sons, Inc., 213-227.
- Yin, R.K. (2003). *Case Study Research: Design and Methods*. Thousand Oaks: Sage Publications.





## Examining survey participation and response quality: The significance of topic salience and incentives

Lazarus Adua and Jeff S. Sharp <sup>1</sup>

### Abstract

Nonresponse bias has been a long-standing issue in survey research (Brehm 1993; Dillman, Eltinge, Groves and Little 2002), with numerous studies seeking to identify factors that affect both item and unit response. To contribute to the broader goal of minimizing survey nonresponse, this study considers several factors that can impact survey nonresponse, using a 2007 Animal Welfare Survey Conducted in Ohio, USA. In particular, the paper examines the extent to which topic salience and incentives affect survey participation and item nonresponse, drawing on the leverage-saliency theory (Groves, Singer and Corning 2000). We find that participation in a survey is affected by its subject context (as this exerts either positive or negative leverage on sampled units) and prepaid incentives, which is consistent with the leverage-saliency theory. Our expectations are also confirmed by the finding that item nonresponse, our proxy for response quality, does vary by proximity to agriculture and the environment (residential location, knowledge about how food is grown, and views about the importance of animal welfare). However, the data suggests that item nonresponse does not vary according to whether or not a respondent received incentives.

Key Words: Survey nonresponse; Survey participation; Leverage-saliency; Prepaid incentives; Item nonresponse; Missing data.

### 1. Introduction

Nonresponse bias has been a long-standing issue in survey research, as it affects all survey research regardless of mode (Nathan 2001). As a result, numerous studies have sought to identify factors that affect both item and unit response/nonresponse in various survey modes (Grove 2006; Trussell and Lavrakas 2004; Davern, Rockwood, Sherrod and Campbell 2003; Teitler, Reichman and Sprachman 2003; Singer, Van Hoewyk and Maher 2000; Singer, Van Hoewyk, Maher 1998; James and Bolstein 1992). While these studies have generated insightful and useful information about the factors that affect survey participation, questions about survey response still remain pertinent to the field of survey research in general and to our substantive work in particular. We are interested in expanding on the thoughts of Groves *et al.* (2000) by investigating whether specific characteristics of sampled units or demographic subpopulations in relation to a survey's topical context affect the response patterns. In our ongoing research assessing the general public's attitudes and behaviours related to the agricultural and environmental domain, we have become increasingly concerned about the level of survey participation and item nonresponse in distinct subpopulations. In our case, one concern is that unit and item nonresponse may vary among individuals or households that are more or less physically or socially proximate to the agricultural landscape, which is the focal area of our public opinion surveys.

To contribute to the broader goal of minimizing item and unit nonresponse and address some of our concerns, we reconsider several factors that can impact survey participation and item nonresponse. Specifically, we examine the effects of a survey's subject context (that is, its main focus) on survey participation and item nonresponse. We anticipate that participation in a survey will be systematically affected by how salient the survey's topic is to each sampled unit. This expectation draws on the leverage-saliency theory (Groves *et al.* 2000), which anticipates that a variety of factors related to a survey's main features or features made prominent during survey administration might impact participation. Our research will also reconsider the effects of prepaid incentives on survey response. Given that offering incentives to sampled units has remained an enduring and widespread practice in the survey industry, we think it behoves survey researchers to periodically reassess the relationship between incentives and survey participation, using varying contexts. Such a continuous assessment of the utility of incorporating incentives into surveys is important because we cannot assume that incentives will always work as intended.

In the next section, we briefly describe the problem of survey nonresponse and then review research on how increasing the salience of some survey features and offering prepaid incentives affect participation and item nonresponse. The final two sections will cover the research design and results of the study.

1. Lazarus Adua, The Ohio State University, 330 Agricultural Admin Building, 2120 Fyffe Road, Columbus, OH43210, U.S.A. E-mail: [adua.1@buckeyemail.osu.edu](mailto:adua.1@buckeyemail.osu.edu); Jeff S. Sharp, The Ohio State University, 254 Agricultural Admin. Building, 2120 Fyffe Road, Columbus, OH43210, U.S.A.

## 2. Survey nonresponse and potential consequences

Survey nonresponse describes the situation in which a sampled unit fails either to participate in the survey altogether (unit nonresponse) or to respond to one or more survey items (item nonresponse). Survey nonresponse has been a long-standing issue in survey research. Singer (2006) observes that “analysis of JSTOR statistical journals dates the first nonresponse article from 1945 and the *Public Opinion Quarterly* index’s earliest reference is from 1948” (page 637). However, well-established and nascent survey projects alike are experiencing steadily declining response rates despite this awareness. For example, the University of Michigan’s Survey of Consumer Attitudes (SCA) has witnessed a drop in response rate from about 72 percent in 1979 to about 60 percent in 1996 and a low of 48 percent in 2003 (Curtin, Presser and Singer 2005).

Survey nonresponse at both the unit and item levels obviously represents a major challenge to survey research, given its potential for generating nonsampling errors in parameter estimates (Brehm 1993; Dillman *et al.* 2002; Groves and Cooper 1998). For example, nonresponse may lead to biased point estimators, variance inflation for point estimators, and biases in estimators of precision (Dillman *et al.* 2002; Groves and Cooper 1998). Although unit and item nonresponse mean different things conceptually in the survey literature, their effects on a statistical estimate are generally the same (Groves, Fowler, Jr., Couper, Lepkowski, Singer and Tourangeau 2004).

While a number of recent studies suggest that low (unit) response rates may not have serious adverse effects on data quality (Curtin, Presser and Singer 2000; Keeter, Miller, Kohut, Groves and Presser 2000; Visser, Krosnick, Marquette and Curtin 1996), the fact still remains that unit nonresponse can have negative consequences for statistical estimates under certain circumstances. As a result, finding creative ways to increase response rates so that all types of sampled units are represented adequately in the sample remains a key goal in survey research. For item nonresponse, it may be true that advances in post-survey techniques for handling missing data, such as hot-deck and cold-deck imputations, mean imputation, multiple imputation, and multiple imputation and deletion, have made it possible to reduce the challenges this poses. However, the ideal situation and, in fact, a primary goal of survey design and implementation is to minimize item nonresponse to the greatest extent possible. This is because the norm in some fields, especially in microeconomics, is to use only the original data (Cameron and Trivedi 2009).

## 3. Making salient key features of a survey and survey participation

The extent to which a sampled unit views some features of a survey as more or less important affects the respondent’s likelihood of participating in the survey (Groves *et al.* 2000). Groves *et al.* (2000) comment on the interviewing tactics of experienced interviewers, arguing that what interviewers actually do when they tailor their queries or remarks to the concerns of respondents is “to heighten the salience of some features of the request, those they judge will be favorably received by the household” (page 299). Building on Groves and Cooper (1998), Groves *et al.* (2000) propose what they call the *leverage-saliency theory* to explain how sampled units make the decision to participate or decline to participate in a survey. This theory essentially states that there are some attributes (leverage) of a survey that may be viewed negatively or positively by the respondent, and that how these attributes are made salient during the survey request process affects the likelihood of participation. If attributes viewed positively by a sampled unit (positive leverage) are made salient during the survey request, there is a higher chance that the respondent agrees to participate in the survey, all other things being equal. On the other hand, the likelihood of a sampled unit participating in a survey will be hurt if attributes that are viewed negatively by the respondent are made salient during the survey request.

Groves *et al.* (2000) empirically support this theoretical position. They present civic engagement (measured by community involvement) and incentives as leverages on survey participation, successfully showing that both attributes positively affect the likelihood of participation, with the effect of incentives diminishing among sampled units with higher civic engagement. In using civic engagement as a measure of a survey’s leverage on sampled units, Groves *et al.* (2000) observe that leverage is not measured directly. Instead, it may be gleaned from some characteristic(s) of respondents in relation to the survey or its features, which may exert a positive or negative influence on the likelihood of participation. There is also evidence that when survey requests are tailored to the concerns of sampled units or to what they consider to be important, the likelihood of their participation is enhanced (Dillman 2000; Groves and Cooper 1998).

Based on the leverage-saliency theoretical proposition, we expect higher rates of participation from respondents whose characteristics make them more likely to view important attributes (leverage) of a survey positively. Correspondingly, we also expect those whose characteristics make them less likely to view such attributes positively to participate in the survey at lower rates. In our particular area



of research, we anticipate that sampled units' proximity to the agricultural and rural landscape (the contextual focus of our on-going survey) will affect participation in the survey and item nonresponse. This logic also applies to our expectations about respondents who claim greater knowledge of how food is produced and who also view animal welfare as important (a central sub-theme of this particular work). We thus draw from the leverage-salience theoretical proposition to propose the following hypotheses.

1. Our survey's focus on agriculture and the environment, which was made salient in its design, is expected to exert a positive leverage on respondents with greater social and physical proximity to agriculture and the rural environment (that is, those residing in more rural places). We thus hypothesize that participation rates will vary according to residential location.
2. We expect respondents with a closer proximity to agriculture and the rural landscape to be more diligent in completing the survey than those not in close proximity, as the former are more likely to be motivated by the survey's subject matter (that is, its positive leverage). We thus hypothesize that item nonresponse will vary by proximity to agriculture and the rural landscape.
3. Sampled units who have greater knowledge of how their food is grown as well as those who view animal welfare as important will have fewer item nonresponses. Presumably, such respondents will have a greater interest in the survey's focus on agriculture and the environment, and therefore exhibit more diligence in completing the survey.

#### 4. Incentives and survey participation

The use of various forms of incentives, particularly prepaid (monetary) incentives, has become a common practice in survey research. While the practical rationale for offering incentives to sampled units is to encourage participation, the theoretical root of this practice is in part traceable to the *social exchange theory* (Dillman 1978). The social exchange theory assumes that people's actions are primarily motivated by the returns they expect or obtain from engaging in an activity (Weisberg 2005). Gouldner (1960) elaborates on the norm of reciprocity, which is related to the social exchange theory, observing that "insofar as men live under such a rule of reciprocity, when one party benefits another, an obligation is generated. The recipient is now *indebted* to the donor, and he remains so until he repays" (page 174). In Gouldner's view, the norm of reciprocity makes two demands on people: (1) people

should help those who have helped them, and (2) people should not injure those who have helped them (Gouldner 1960, page 171).

Dillman (1978) uses the social exchange theory and particularly the social norm of reciprocity to argue that relatively small gestures (such as personalized letters, incentives, and reminder letters) can evoke reciprocation from sampled households in terms of inclination to participate in a survey. Also, Weisberg (2005) notes that social exchange is a theory that possibly explains the relationship between incentives and survey participation, observing that "[f]rom this perspective, giving the respondent a monetary incentive to participate in the survey can be seen as a kindness that evokes a norm of reciprocity" (page 165).

To devise 'ways and means' to bolster survey response rates as well as to test the social exchange theory in relation to incentive use in survey research, a number of experimental studies have examined the relationship between providing incentives to respondents and survey participation. While some of these studies have focused primarily on the effects of incentives on response rate and item nonresponse (Grove, Couper, Presser, Singer, Tourangeau, Acosta and Nelson 2006; Trussell and Lavrakas 2004; James and Bolstein 1992; Church 1993; Singer 2000; Yammarino, Skinner and Childers 1991; Fox, Crask and Kim 1988), others have examined the effects of incentives on respondent expectations and views about surveys (James and Bolstein 1990; and Singer *et al.* 1998). Consistent with the main proposition of the exchange theory and the norm of reciprocity, many of these studies report a positive relationship between incentives and response rates (Singer *et al.* 2000; Groves, Couper, Presser, Singer, Tourangeau, Acosta and Nelson 2006; Church 1993; Trussell and Lavrakas 2004; Goyder 1982; and Yu and Cooper 1983).

While many studies confirm the importance of incentives in encouraging survey participation, the empirically informed verdict on the relationship between incentives and survey participation is by no means unanimous. In a meta-analysis of experimental and quasi-experimental studies involving incentive conditions, Church (1993) reports that 1% of the studies utilized found no evidence of incentives affecting participation. Church also states that 10% of the 74 studies analyzed actually reported a negative relationship between the incentive conditions and survey participation. In fact, this reality partly prompted Groves *et al.* (2000) to propose the leverage-saliency theory to help explain why "incentives sometimes work" but "sometimes don't" (page 299). Given that findings related to the effects of incentives on survey participation are moderately mixed, as well as the fact that the subject matter of the survey we are studying differs from many previous studies, we find it necessary to

assess incentive effects on survey participation in conjunction with our examination of the relationship between agricultural proximity (our survey's contextual focus) and response. Also, we believe it is important to periodically assess the utility of using incentives in survey research, despite the fact that this subject has received a lot of attention in the past.

Another important incentive-related issue is the potential higher item nonresponse impacts of inducing reluctant respondents to participate in a survey (see Hansen 1980). The potential harm exists in that using persuasions such as incentives might elicit information from respondents who are careless or indifferent when answering questions, ultimately damaging the quality of the information obtained in this way (Singer *et al.* 2000). Owing to this concern, a number of studies have examined the relationship between incentives and item nonresponse, many of which suggest that incentives do not seriously harm response quality; that is, incentives do not generate higher item nonresponse (Singer *et al.* 2000; Singer *et al.* 1998; Shettle and Mooney 1999 and Davern *et al.* 2003). In fact, Singer *et al.* (2000) actually report that prepaid incentives help to reduce item nonresponse, an often-used measure of response or data quality. However, they also report that respondents who received incentives were more likely to give optimistic answers in some cases and be more pessimistic in others (involving different variables). In our case, a critical concern is that urban respondents induced to participate may provide lower quality data (as measured by nonresponse) than respondents more proximate to the agricultural and rural landscape.

In summarizing the review, we find that the research generally suggests that incentives help improve response rates in surveys, with little or no effect on item nonresponse. Although this is generally the case, some findings on the relationship do deviate from this expectation (Church 1993). Also, while many studies find that providing prepaid incentives does not affect item nonresponse, the work of Singer *et al.* (2000) suggests that providing incentives can compromise data quality via the mechanism of optimism or pessimism bias. Given these caveats, as well as the fact that most prior work on the relationship between incentives and survey participation was based on bivariate analysis (incentive and survey participation), we find it necessary to reconsider the impact of incentives on survey nonresponse while taking into account the effects of residential location in space and socioeconomic status. Thus, drawing from this literature on how incentives are related to survey participation and item nonresponse, we make the following hypotheses.

1. Respondents who received incentives will participate in the survey at higher rates than non-recipients, net

the effects of proximity to the agricultural and rural landscape and socioeconomic status.

2. Incentives will be negatively related to item non-response. That is, surveys completed by respondents who received incentives will have fewer missing data points than those completed by respondents who did not receive incentives, controlling for the effects of respondents' proximity to the survey's subject and other covariates.

## 5. Study design

This paper is based on a survey of public views regarding food, agricultural and environmental issues, with a special focus on farm animal welfare. The target population of the survey was Ohio households. An initial sample of 3,000 respondents (along with their residential addresses) was drawn for the study via stratified random sampling: one-half (1,500) from Ohio's 22 core metropolitan counties and the second half (1,500) from the state's 66 metropolitan fringe or non metropolitan counties. The number of households in the core metropolitan counties differed from those in the metropolitan fringe or non metropolitan counties, making the sample a disproportionate random sample. To account for the unequal probability of selection across the two strata, we conducted weighted analysis for this paper.

The sample we used was obtained from Experian, a U.S.-based credit reporting bureau and private list vender. The sample was drawn from a sample frame (database) consisting of Ohio households along with their residential addresses. While we do not pretend that this sample frame covers all Ohio households, we believe that it is one of the most reliable and up-to-date lists and databases in the U.S. from which one can draw a sample. According to Experian, the database is updated monthly.

The survey followed a modified tailored design method (Dillman 2000) with up to four mailings sent to potential respondents during the spring of 2007. The first mailing was a pre-notification letter sent to each sampled unit, followed shortly by the survey packages. The third mailing was a reminder postcard sent to respondents thanking them for participating in the study or encouraging them to complete and return the survey if they had not yet done so. In the fourth mailing, replacement survey packages were mailed to respondents who had not returned completed questionnaires about 10 days after the postcard was mailed out. Of these four contacts with the respondents, three had information that focused specifically on the subject or topic of the survey. The pre-notification letter and the cover letters for the initial and replacement survey packages specifically conveyed to respondents the subject matter of the survey. Also, the graphics printed on the cover page of the survey



(images of farm animals) were selected to further convey this subject matter.

The addresses of sampled units were geo-coded and placed in a locational field (see details later in this section) to locate them geographically across the rural-urban continuum. This allowed us to conduct analyses of how sampled units' proximity to the agricultural landscape is related to their likelihood of participating in the survey. We recognize that some urban residents may have frequent social and physical interactions with agriculture and the rural landscape; however, this kind of interaction, along with its effects on support for agriculture and the environment, is highest among those residing in more rural and open country places (Freudenburg 1991; Sharp and Adua 2009). A randomized experiment involving incentives was also built into the survey. The first survey packages mailed to a randomly-selected half of the sampled units included \$2.00 (two one dollar bills) incentives, while the other half of the sample received the same package but without any incentives. In doing this experiment, our pragmatic objective was to assess the effectiveness of our practice of enclosing modest cash incentives in survey packages to improve participation in our ongoing surveys of the Ohio public. Similar to Groves *et al.*'s (2000) expectations about the effect of community involvement on levels of participation, we also anticipated that households located in close proximity to agriculture and the rural landscape would participate at high levels in our study independent of the incentive, perhaps to the extent that a token financial incentive might be deemed unnecessary in future iterations of the survey.

### 5.1 Analytic strategy

Two sets of statistical analyses are conducted in this paper. The first set of analyses focuses on survey participation (response rate). First, we examine the proportion of successfully contacted sampled units who complete and return surveys by residential location along the rural-urban continuum, a proxy for geographic proximity to agriculture and rural areas of the state (an assumption we justify in a later section), and by incentive status. Following the American Association of Public Opinion Research's (AAPOR) 2008 guidelines for codes disposition, we defined successfully contacted sampled units as (i) those from whom we received completed surveys by the end of the data collection phase of the project, and (ii) those from whom we received neither a completed survey nor the survey package back from the United States Postal Service (USPS) as undeliverable. In our contract with the USPS, we requested that all mails that could not be delivered due to wrong address or absence of forwarding information be returned to us. The sampled units to which these undeliverable mails

were addressed were classified as units we were unsuccessful in contacting. We also employ logistic regression to further analyze the likelihood of survey participation (coded 1 = responded; 0 = did not respond), using residential location along the rural-urban continuum and incentive status as the primary predictors, while simultaneously controlling for the effects of socioeconomic status at respondents' block group level as per the 2000 U.S. population census. We control for the effect of socioeconomic status because previous studies suggest it has some relationship with survey participation (Davern *et al.* 2003; Singer *et al.* 2000).

The second set of analyses focuses on item nonresponse. In this analysis, we conduct partial proportional ordered logistic regression analysis (generalized ordered logit) on the first two item nonresponse variables (0 = no missing items; 1 = some missing items; and 2 = numerous missing items), once again employing residential location along the rural-urban continuum and incentive status as the primary independent variables while controlling for the effects of several other variables. Generalized ordered logit (partial proportional odds) is employed rather than ordered logit because some predictors in these models violated the proportional odds assumption of ordered logistic regression. By using partial proportional odds modeling, we are able to constrain the relationship between those independent and dependent variables that met the proportional odds assumption of ordered logistic regression while allowing the relationships that failed this assumption to vary. To analyze the third item nonresponse variables, we employed logistic regression. This variable was recoded into a dichotomy (see the section on operationalization of variables for more details).

### 5.2 Operationalizing dependent variables

*Survey Participation:* Survey participation (response rate) is measured by computing the number of completed surveys received from respondents (eligible participating cases) as a proportion of the sampled units contacted successfully (all eligible cases). This measure of survey participation is in conformity with AAPOR guidelines for measuring response rates. Undeliverable surveys returned by the USPS without additional information, such as forwarding address or address correction, were treated as ineligible. Cases for which we neither received completed surveys nor any other information about the cases from the USPS were treated as eligible based on the recommendation of the AAPOR's 2008 revised standard definitions of codes disposition and outcome rates. To conduct the logistic regression analysis of response likelihood, we coded all successfully contacted sampled units (eligible cases) as 1 (returned a completed questionnaire) or 0 (did not return a



completed questionnaire). We provide no descriptive statistics for this variable here as the analysis section, especially the marginals of the contingency tables, provides a good sense of the distribution of this variable.

**Response quality:** Response quality is measured by the occurrence of item nonresponse (see Davern *et al.* 2003; and Kaldenberg, Koenig and Becker 1994). To compute item nonresponse, missing data points for all respondents participating in the survey were summed across three subsets of items in the survey instrument to generate three item nonresponse variables: item nonresponse I, item nonresponse II and item nonresponse III. The item nonresponse I variable was created from items that, in our estimation, exerted comparatively the lowest cognitive demand on respondents, including such items as demographics and opinion questions that did not require very much introspection. The item nonresponse II variable was created from items that exerted comparatively higher cognitive demands on respondents than those used to create item nonresponse I, such as questions that required significant recall efforts and opinion questions that required a high level of introspection. The third variable is constructed from items that exerted comparatively the highest cognitive demand on respondents, such as knowledge questions and questions that required some understanding of concepts associated with animal husbandry.

In summing across these variables, we did not treat 'Don't Know' answers as item nonresponse, given that the survey had a couple of knowledge questions for which a 'Don't Know' response could be a legitimate answer. The item nonresponse variable also does not include "refused to answer" responses, as this option was not provided in questions used in the creation of the variables. We also excluded from these variables questions that respondents were directed to skip if they found them to be inapplicable.

Owing to the fact that the distribution of these variables was heavily skewed (see Table 1), the item nonresponse I and nonresponse II variables were regrouped into three ordinal categories (0 = no missing items; 1 = some missing items; and 2 = numerous missing items) and analyzed using generalized ordered logit. The first category (0) included cases without any item nonresponse, while the second category (1) included cases with between 1 and 9 incidences of nonresponse. The third category (2) included cases with 10 or more item nonresponses. For our analysis, we also regrouped the item nonresponse III variable into a dichotomy: 0 (no missing cases) and 1 (1 or more missing cases). This variable was regrouped differently from the first two because very few cases (only 19) satisfied the criteria for classification as "numerous missing cases" (Table 1). To verify whether our regrouping of these variables masked variances in item nonresponse within the groups (cases

grouped together) that may be explained by our two independent variables (residential location, i.e. an indicator of interest in the survey topic, and incentives), we conducted a one-way analysis of variance for these grouped cases. Within these groups, none of the three item nonresponse variables varied significantly by residential location or incentives. Descriptive statistics for all three item nonresponse variables are reported in Table 1.

**Table 1**  
Descriptive statistics for item nonresponse variables

	Item nonresponse I	Item nonresponse II	Item nonresponse III
<i>Statistics before recoding</i>			
N	971	971	971
Mean	3.11	2.34	1.6
Standard deviation	5.06	5.93	3.25
Minimum value	0	0	0
Maximum value	44	48	29
<i>Statistics after recoding into groups</i>			
Zero missing	30.07%	59.53%	54.69%
Some missing	62.31%	32.65%	43.36%
Numerous missing	7.62%	7.83%	1.96%

### 5.3 Operationalizing independent and control variables

**Residential Location:** The survey's focus on agricultural and environmental issues was made salient during the survey request (via the pre-notification letters, the cover letters and the design of the survey instrument), which can affect participation negatively or positively depending on each respondent's residential location along the rural-urban continuum. Residential location is an indicator of respondents' differentiated social and physical proximity to agriculture and the rural landscape. This is because proximity can increase the social and/or physical interactions with the subject. The association between proximity and environmental concern has been proposed and tested numerous times by social scientists (Dunlap and Heffernan 1975; Freudenburg 1991; Sharp and Adua 2009). We go a step beyond hypothesizing attitudinal differences associated with proximity and anticipate different levels of survey participation; indeed, we hypothesize that sampled units residing closer to agriculture and the rural landscape will participate in the survey at higher rates than those in core urban places. As a result, the subject matter of our survey is expected to serve as a positive leverage on sampled units residing closer to agriculture and the rural landscape. While this may not be a direct measure of leverage, it is consistent with Groves *et al.*'s (2000) suggestion that the leverage a given survey exerts on a sampled unit can be measured indirectly by relying on pertinent characteristics of the

sampled unit. In using the spatial residential characteristics of sampled units, we are relying on the fact that sampled units residing in more rural and open country areas have a higher likelihood of social and physical interaction with the agricultural and rural landscape than those in more urbanized places (see Table 2). In both 2006 and 2007, higher proportions of residents of exurban townships and rural areas (a combination of rural city/village and rural townships) visited a working farm than residents of core urban places, as shown in Table 2. We acknowledge that using information from our own respondents to show the association between residential location and visits to farms may be problematic. However, this information is corroborated by information from a different sample, the 2006 Ohio Survey.

To determine the residential location of the sampled units, each respondent's residential address was geocoded and assigned to one of four location fields—urban, suburban, exurban or rural—using ESRI's ArcView geocoding. Sampled units living in the exurban and rural fields were further distinguished as residing in either incorporated places (city/village) or township places (open country). This process of characterizing sampled units as living in urban, suburban, exurban, or rural places has previously been employed successfully in the field of regional science (Audirac 1999; Sharp and Clark 2008).

In this study, this variable has been grouped into five categories: (1) core urban, (2) suburban places, (3) exurban city/village, (4) exurban township and (5) rural places (cities/villages and townships). The ordering of the categories does not suggest a monotonic increasing order in terms of proximity to agriculture and the rural landscape between categories 1 and 5. Instead, this variable should be seen as a nominal variable with categories that can be grouped into blocks based on proximity to agriculture and the rural landscape: block 1(categories 1 and 2) has the lowest proximity, block 2 (category 3) has intermediate proximity and block 3 (categories 4 and 5) has the highest proximity.

Between the blocks, the categories are monotonic increasing in terms of proximity to agriculture and the rural landscape, but within the blocks the pattern is less certain. Here, too, we provide no descriptive statistics for this variable as the analysis section provides an ample sense of how the variable is distributed.

*Knowledge of Food Production and Support for Animal Welfare:* Two other indicators of survey leverage used in the analysis are two survey items that measured sampled units' knowledge of how their food is produced and their views about the importance of animal welfare. The first asked, "How knowledgeable are you about how your food is grown? Please indicate on a scale of 1 to 7 your level of knowledge." This item had a mean of 4.47 and a standard deviation of 1.60. The second item asked, "Thinking about farm animals in general, how important is this issue to you? Please indicate on a scale of 1 (not important) to 7 (very important)." This item had a mean score of 4.50 and a standard deviation of 1.68. These two indicators are used in analyses pertaining only to the item nonresponse variables.

*Incentive Status:* Sampled units' incentive status (received versus did not receive incentive) is a primary independent variable in the regression models. Incentive status is dummy-coded as 0 (did not receive incentive) and 1 (received incentives). Again, we provide no descriptive statistics for this variable because the analysis provides a good sense of the variable's distribution.

*Control Variables:* Control variables operationalized in one or more of the analysis conducted in this study include *Age* (respondent's age as of his/her last birthday), *Education* (highest level of education completed), *Ethnicity* (white = 1; all others = 0) and *Gender* (male = 0 and female = 1), as well as the per capita and disposable median household income of each sampled unit's block group as per the 2000 population census. We control for the effects of these variables because previous studies suggest they can affect item nonresponse (Davern *et al.* 2003; Singer *et al.* 2000). Descriptive statistics for these purely control variables are shown in Table 3.

**Table 2**  
**Frequency of visiting or touring a working farm**

Residential location	2006 Ohio Survey <sup>a</sup>			2007 Animal Welfare Survey <sup>b</sup>		
	Never/ seldom	Occasional/ frequently	Total <sup>c</sup>	Never/ seldom	Occasionally/ frequently	Total <sup>c</sup>
Core urban	90.4%	9.6%	100% (185)	81.0%	19.0%	100% (121)
Suburban place	87.5%	12.5%	100% (536)	83.7%	16.3%	100% (285)
Exurban city/village (Incorporated)	78.6%	21.4%	100% (217)	76.4%	23.6%	100% (124)
Exurban township (Unincorporated)	74.9%	25.1%	100% (434)	67.9%	32.1%	100% (264)
Rural place	73.1%	26.9%	100% (238)	70.6%	29.4%	100% (136)
Total	80.6%	19.4%	100% (1,610)	74.2%	25.8%	100% (930)

<sup>a</sup> Second-order corrected chi-square (3.61) = 43.3;  $P = 0.0000$  (corrected for survey design effects)

<sup>b</sup> Second-order corrected chi-square (3.67) = 16.7;  $P = 0.001$  (corrected for survey design effects)

<sup>c</sup> In parentheses are the total number of eligible cases from each residential category.



**Table 3**  
**Descriptive statistics for control variables**

	Mean/percent	Standard deviation
<i>Education:</i>		
High school and lower	36.8%	-
Some college	32.3%	-
Bachelor's degree	13.7%	-
Grad/professional work & higher	17.2%	-
<i>Gender:</i>		
Male	48.2%	-
Female	51.8%	-
<i>Ethnicity:</i>		
White	91.7%	-
Non-white	8.3%	-
<i>Age:</i>	51.9	15.8
Block level mean household income, 2000	49,842.3	25,258.7
Block level median household income, 2000	42,616.3	16,728.6

## 6. Results

To evaluate survey participation, we use both bivariate analysis (contingency tables) and logistic regression modeling. For the contingency tables, we use Pearson chi-squared statistics corrected for survey design with Rao and Scott's (1984) second-order correction. We do this because survey design features such as stratification and clustering can affect tests of association (Lohr 1999). To limit the length of this paper, we follow a different analytical plan for the item nonresponse set of variables. For this set, we conduct only multivariate analysis (logistic regression). Moving straight to multivariate analysis allows us to examine the partial effects of the various predictors used in the models while keeping the paper brief.

### 6.1 Bivariate results for survey participation

The bivariate analysis suggests that survey participation varies significantly by proximity to the agricultural and rural landscape (residential location along the rural-urban continuum). As shown in Table 4, respondents residing in geographically more rural places (rural and exurban township residents) have higher rates of participating in the survey than those residing in geographically more urban places (core urban and suburban residents). The analysis also shows that those in the intermediate exurban incorporated places (cities and villages) were slightly more likely to participate than core urban residents. A second-order corrected chi-square test (Rao and Scott 1984) of the relationship between survey participation and residential location was significant ( $\chi = 14.2$ ;  $df = 3.7$ ; and  $p = 0.003$ ).

Our analysis is consistent with previous studies, also finding that prepaid incentives significantly increase survey

participation (Table 5). Despite the fact that the context of the survey used for our analysis differs markedly from previous studies examining the effects of incentives, we find that the response rate for successfully contacted incentive recipients was 43.7% compared with 26.9% for successfully contacted sampled units who did not receive prepaid incentives. The second-order corrected chi-square test of this bivariate relationship is also statistically significant ( $\chi = 73.8$ ;  $df = 1$ ;  $p = 0.000$ ). In fact, our analysis suggests that eliminating incentives altogether substantially hurts participation rates for all categories of respondents regardless of proximity to the agricultural and rural landscape, although this effect is highest for residents in core urban places (Table 6). This finding provides support for our ongoing practice of using prepaid monetary incentives to help bolster our response rates with no discrimination between whether respondents reside in rural or urban locales. It also reaffirms the importance of incentives in survey research.

### 6.2 Logistic regression model for survey participation

Multivariate analysis further suggests that the likelihood of survey participation varies significantly by proximity to agriculture and the rural landscape, statistically holding constant the effects of incentive status (received versus did not receive incentive). Residents of suburban places, exurban townships, and rural places are significantly more likely to participate in the survey than residents of core urban places (Table 7). For example, residents of exurban townships and rural places have higher odds (0.60 log odds and 0.37 log odds, respectively) of participating than those of core urban places.



**Table 4**  
**Participation rate by residential location**

Residential location	Responded	Did not respond	Total <sup>a</sup>
Core urban	29.5%	70.5%	100% (424)
Suburban place	32.6%	67.4%	100% (917)
Exurban city/village (Incorporated)	33.1%	66.9%	100% (379)
Exurban township (Unincorporated)	40.5%	59.5%	100% (684)
Rural place	35.8%	64.2%	100% (405)
Total	35.4%	65.6%	100% (2,809)

Second-order corrected chi-square (3.7) = 14.2;  $P = 0.003$  (corrected for survey design effects)

<sup>a</sup> In parentheses are the total number of eligible cases from each residential category

**Table 5**  
**Survey response by incentive status**

Incentive status	Responded	Did not respond	Total <sup>a</sup>
Incentive	43.7%	56.3%	100% (1,410)
No incentive	26.9%	73.1%	100% (1,401)
Total	35.4%	64.6%	100% (2,811)

Second-order corrected chi-square (1) = 73.8;  $P = 0.000$  (corrected for survey design effects)

<sup>a</sup> In parentheses are the total number of eligible cases by incentive status

**Table 6**  
**Response rate by incentives and residential location along the rural-urban continuum**

	Incentive recipients	Non-recipients of incentive	Response difference
Core urban	0.41	0.19	0.22
Suburban place	0.41	0.24	0.17
Exurban city/village (Incorporated)	0.39	0.27	0.12
Exurban township (Unincorporated)	0.48	0.31	0.17
Rural place	0.44	0.27	0.17
Total	0.43	0.26	0.17

Logistic regression analysis also seems to confirm our earlier finding that the likelihood of participating varies significantly by whether or not a sampled unit received incentives. Respondents who received incentives had higher odds (0.73 log odds) of participating in the survey than those who did not receive incentives, controlling for proximity to agriculture and the rural landscape as well as the gender (female = 1) of the householder randomly assigned as the preferred household member to complete and return the survey (Table 7).

Because socioeconomic status varies significantly by residential location across space (Lobao 1990) and affects survey response (Davern *et al.* 2003; Singer *et al.* 2000), we endeavored to control for the potential effects of per capita income and household income (socioeconomic status) on the likelihood of survey participation using hierarchical linear modeling (HLM). To do this, respondents were linked to their block groups and block group characteristics

(specifically, block group per capita income and block group household median income) as per the 2000 U.S. population census. For the HLM analysis, we initially estimated a fully unconditional model (that is, an ANOVA) to determine whether the likelihood of survey participation varied significantly across the block groups. In hierarchical linear modeling, estimating a fully unconditional model (model without predictors at all levels of the analysis) is typically used to determine whether the dependent variable varies by the level two (or higher) unit of analysis, such as a neighborhood, block group or school district. This initial model (ANOVA) often helps researchers determine whether to proceed with multi-level analysis. Our initial HLM analysis (ANOVA) did not reveal any significant variation in the likelihood of survey participation across the block groups ( $\tau = 0.04$ ;  $p = 0.493$ ). While this finding suggests the average probability of survey participation is about the same for all block groups despite their different per capita

and household disposable median incomes, we acknowledge potential instability in this HLM model given that sample cases per block group were generally low. This may have led to our finding of no significant variation in the likelihood of participation across the block group (potential Type II error). Despite this potential problem with our fully unconditional model, we did not proceed with the fully conditional multi-level analysis.

### 6.3 Logistic regression model for item nonresponse

As noted earlier in this section, our analysis of item nonresponse is limited to multivariate modeling, and we do this primarily to keep the paper brief while achieving our objective of assessing the partial effects of our main independent variables. The data suggest that the anticipated leverage of the survey's subject is only modestly related to item nonresponse. With respect to item nonresponse I (that is, the variables created from questions with the least cognitive demand on respondents in the survey), the analysis suggests that respondents in exurban township areas have lower item nonresponse (-0.74 log odds) than those residing in core urban areas, although this difference disappears at the higher values of this variable (Table 8, Columns 2 and 3). However, for item nonresponse II (the item nonresponse variables created from questions more cognitively demanding than those used in item nonresponse I) we find that residents of exurban townships and rural places are more likely to have higher item nonresponses (0.85 and 0.82 log odds, respectively) than residents of core urban areas (Table 8 Column 4). In terms of item nonresponse III (the item nonresponse variables created

from the most cognitively demanding questions), the analysis did not reveal any significant difference by residential location, our proxy for level of interest in the survey's topic.

Supporting the anticipated effect of interest in a survey's topic on item nonresponse, the analysis also suggests that respondents' knowledge of how food is produced is significantly related to item nonresponse. In terms of item nonresponse II, the data shows that respondents who reported knowing how food is produced have lower log odds (-0.13) of item nonresponse than those who reported having less knowledge of how food is produced (Table 8, Column 4). This relationship is stronger at higher values of the variable: knowledge of how food is produced has lower log odds (-0.35) of item nonresponse when the category value shifts from 0 to 1 (Table 8 Column 5). This result suggests that the positive leverage of the survey's topic may have resulted in greater care in the completion of the survey among respondents with greater knowledge of how food is produced. We also find that respondents' views about the importance of animal welfare, a central subtheme of this particular survey, are positively related to item nonresponse (Table 8, Column 4). As shown in Table 8, a one unit increase in viewing animal welfare as important leads to a 0.09 unit increase in the log odds of item nonresponse (specifically item nonresponse II). This finding is inconsistent with our expectations.

In terms of the effects of incentives, we find no significant relationship between incentives and any of the three measures of item nonresponse (Table 8, Columns 2, 4 and 6), contrary to our expectation.

**Table 7**  
**Logistic regression<sup>a</sup> of likelihood of participation**

	Log odds of participation	
	b	Std. Error
<i>Incentive status</i>		
Did not receive incentive (Ref)	-	-
Received incentive	0.73***	0.09
<i>Residential location</i>		
Core urban residents (Ref)	-	-
Suburban residents	0.27*	0.13
Exurban city/village residents	0.25	0.15
Exurban township residents	0.60***	0.13
Rural residents	0.37*	0.15
First option to respond (Female = 1)	-0.05	0.09
<i>Model statistics</i>		
Intercept	-1.42***	
Wald $\chi$ (df = 6)	93.25***	

Significance: \*\*\*< 0.001; \*\*< 0.01; and \*< 0.05

<sup>a</sup> In this model we tested for potential interaction effects between residential location and incentives. We found no evidence of such an interaction effect.

**Table 8**  
**Logistic regression models<sup>a</sup> for item nonresponse**

	Item nonresponse I <sup>b</sup>		Item nonresponse II <sup>b</sup>		Item nonresponse III <sup>c</sup>
	No missing: log odds	Some missing: log odds	No missing: log odds	Some missing: log odds	Log odds
<i>Incentive status</i>					
Did not receive incentive	-				
Received incentive	0.16 (0.16)		0.10 (0.17)		-0.01 (0.17)
<i>Subject salience –Residential location</i>					
Core urban residents	-				
Suburban residents	-0.14 (0.26)		0.54 (0.29)		-0.18 (0.25)
Exurban city/village residents	-0.36 (0.31)		0.30 (0.34)		-0.24 (0.29)
Exurban township residents	-0.74** (0.27)	0.30 (0.40)	0.85** (0.30)		-0.12 (0.26)
Residents of rural places	-0.21 (0.3)		0.82** (0.31)		0.08 (0.29)
<i>Subject salience –Food knowledge and animal welfare</i>					
Knowledge about how food is produced	-0.07 (0.05)		-0.13* (0.05)	-0.35*** (0.09)	-0.02 (0.06)
Importance of animal welfare	0.10 (0.05)		0.09* (0.04)		0.10 (0.05)
<i>Controls</i>					
Education:					
High school and lower					
Some college	-0.79*** (0.20)		0.13 (0.19)		0.07 (0.19)
Bachelor's degree	-1.08*** (0.23)		-0.32 (0.27)		-0.52 (0.29)
Grad/professional work & higher	-0.99*** (0.24)		0.12 (0.24)		-0.38 (0.24)
Age	0.03*** (0.01)		0.04*** (0.00)		0.03*** (0.01)
Gender (Female = 1)	0.03 (0.17)		0.53** (0.17)		0.21 (0.17)
White	-0.38 (0.32)		-0.05 (0.28)		-0.51 (0.32)
<i>Model statistics</i>					
Intercept	0.16	-4.36	-3.56	-5.07	-3.07
Wald chi-square <sup>d</sup>		85.80		93.25	54.87
N		828		828	828

Significance: \*\*\*< 0.001; \*\*< 0.01; and \*< 0.05

Standard errors shown in parentheses.

<sup>a</sup> We tested for potential interaction effects between residential location and incentives, between age and incentives and between ethnicity (white) and incentives in these models following Singer *et al.* (2000). We found no evidence of such interaction effects.

<sup>b</sup> The item nonresponse I and II models are partially constrained proportional odds logit models. This is because some of the predictors of these models violated the parallel lines assumption. These predictors were thus allowed to vary, while the remaining ones were constrained. William's (2006) gologit2 stata program code was used to estimate the model.

<sup>c</sup> This model is a logistic regression model with a binary dependent variable (variable recoded into two categories).

<sup>d</sup> Degrees of freedom are 14, 14, and 13 for the low cognitive, mid cognitive, and high cognitive models, respectively.



In terms of the control variables, we find that education is significantly related to item nonresponse, which is consistent with the earlier findings of Singer *et al.* (2000). In our case, respondents with some college work, a bachelor's degree, or some graduate/professional work have lower odds (-0.79, -1.08, and -0.99 log odds respectively) of missing cases for the survey questions with the lowest cognitive demand (item nonresponse I) than those with only a high school education or less (Table 8, Column 2). Surprisingly, item nonresponse related to the survey questions that were comparatively higher in cognitive demand (that is, item nonresponse II and item nonresponse III) did not differ by education (Table 8, Columns 4 and 6). We also find positive relationships between age and all three measures of item nonresponse (Table 8, Columns 2, 4, and 6), which is consistent with Singer *et al.* (2000). Equally consistent with the earlier work of Singer *et al.* (2000), the analysis reveals that female respondents are more likely to have missing data points than male respondents (Table 8, Column 4). However, the effect of gender on item nonresponse in our study is limited to those survey questions with a medium level of cognitive demand (the item nonresponse II variable).

## 7. Discussion and conclusions

In this study, we examined factors related to both unit and item nonresponse in survey research, focusing on interest in a survey's topic and prepaid incentives. The obvious reason for carrying out this analysis is the fact that nonresponse (unit or item) represents a major challenge to survey research given its potential for generating non-sampling errors in parameter estimates (Brehm 1993; Dillman *et al.* 2002; Groves and Cooper 1998). As previously noted, nonresponse can lead to biased point estimators, variance inflation for point estimators, and biases in estimators of precision (Dillman *et al.* 2002; Groves and Cooper 1998). Therefore, our primary goal is to provide information that will help researchers understand and deal appropriately with nonresponse, that is, minimize unit nonresponse and correctly understand and handle missing cases (item nonresponse).

Our analysis reveals that the likelihood of participation in this survey on agriculture and the environment varies significantly by sampled units' proximity to the agricultural and rural landscape (residential location). Our analysis is consistent with our first hypothesis and the theoretical proposition of leverage-saliency, as we find that residents of exurban townships and rural places are all significantly more likely to participate in the survey than residents of core urban places. The pattern of relationships revealed in this

analysis is most likely explained by the fact that respondents residing in exurban townships and rural places have a higher chance of interacting with the agricultural and rural landscape than those residing in core urban places (see Table 2). Thus, we suggest that respondents residing closer to the agricultural and rural landscape participated at higher rates in the survey due to the positive leverage of the survey's focus on the agricultural and environmental domain.

We also find some relationship between interest in the survey's topic (measured by proximity to the agricultural and rural landscape) and response quality (measured by item nonresponse). In support of our second hypothesis, modest evidence in this study suggests that item nonresponse varies by proximity to the agricultural and rural landscape. For item nonresponse I, the data suggest that residents of exurban township areas are less likely to have missing data points than residents of core urban places, whereas residents of both exurban townships and rural places are more likely to have missing data points for item nonresponse II. Missing cases associated with questions with the highest cognitive demand (item nonresponse III) did not vary by residential location (interest in the survey's topic). These findings suggest that residents of the more rural places (exurban townships and rural places) fare worse than those of core urban places when missing cases involve survey questions with a moderate level of cognitive demand. Although this result is intriguing, we are unable to explain why it is the case. One possible argument would be the educational difference between residents of core urban and rural places, but this study statistically controls for the effects of education. Further work certainly needs to be done on this subject.

Knowledge of how food is produced, another indicator of proximity to agriculture and the rural landscape, is negatively related to item nonresponse, which is consistent with our expectation (hypothesis 3) and the leverage-saliency theory. As the knowledge of how food is produced is related to the broader topic of the survey, we believe that making the survey's focus on agriculture and the environment salient in our request for participation in the survey may have generated higher diligence in questionnaire completion among respondents who knew or cared enough to know how food is produced. However, our analysis also suggests that support for animal welfare is positively related to item nonresponse, which is inconsistent with hypothesis 3. These findings highlight the need to look closely at factors related to a survey's topic as potential covariates of item nonresponse and its corollary, nonresponse error.

Although the survey used in this study focused on agriculture and the environment, our findings in relation to the survey's topic may have implications for surveys that focus on other sectors. There is reason to believe that unit

and item nonresponse can be affected by respondents' proximity to or level of interest in any survey topic or industry of focus, especially if this aspect of the survey is made salient during the request for participation. For example, if a survey focuses on the automotive industry and this feature is made salient during the request for participation, it is very likely that this information will affect the response pattern. In essence, these findings suggest that researchers designing surveys need to think critically about how the survey's subject context, such as the industry or sector on which it focuses, might affect participation from subpopulations within the sample list. While this generalization may be reasonable, we believe similar studies focusing on other sectors will be required before we can draw firm conclusions.

We next discuss the relationship between prepaid incentives on the one hand and survey participation and item nonresponse on the other. With respect to the relationship between incentives and response, our study suggests that prepaid incentives generally increase the likelihood of a respondent participating in a survey, even if proximity to agriculture and the rural landscape (the survey subject context) is taken into account. Our findings are consistent with hypothesis four and the previous literature (Singer *et al.* 2000; Groves 2006; Church 1993; Trussell and Lavrakas 2004; Goyder 1982; and Yu and Cooper 1983), as they show that recipients of prepaid incentives were significantly more likely to participate in the survey than non-recipients, controlling for other variables in the logistic regression model. The analysis demonstrates that eliminating incentives altogether hurts the likelihood of participation regardless of respondents' residential context. While we may not have overtly identified prepaid incentives with the leverage-saliency theory of Groves *et al.* (2000) in the earlier sections of our discussion for the sake of analytical convenience, our findings in relation to this variable also provide further empirical support for this theory. Our findings clearly suggest that token financial incentives enclosed with each survey package helped increase participation from both metropolitan and non-metropolitan areas of Ohio, although this effect was higher in the former. This result provides fresh justification for the widespread use of incentives to bolster response rates. As indicated earlier in this paper, the widespread use of prepaid incentives in surveys makes it necessary to periodically assess the utility of this practice. Our finding also suggests the need to check for potential response bias if incentives are provided to only a section of the sampled respondents, such as when prepaid incentives are targeted at those assessed as being less likely to participate.

In terms of the relationship between incentives and item nonresponse, we find no significant variation in missing

data points between respondents who received monetary incentives and those who did not, contrary to our fifth hypothesis. This finding, which controls for the effects of residential location (proximity to the agricultural and rural landscape) and other pertinent variables, is consistent with the earlier work of Davern *et al.* (2003), who failed to find any relationship between incentives and the number of imputations for missing data points. Thus, while the use of monetary incentives correlates significantly with unit nonresponse (outright nonparticipation in a survey), we find no relationship between incentives and item nonresponse (failure to respond to some questions on a questionnaire). Thus, providing incentives to a respondent does not necessarily lead to greater diligence in survey completion.

The analysis revealed some interesting results with respect to the relationship between some of the control variables and item nonresponse. While education, age and gender were used in this study primarily as control variables, the fact that they were found to be significantly related to item nonresponse raises practical concerns about handling missing cases in survey data. Before choosing between the various techniques for handling missing cases (see Fuchs and Kenett 2007), analysts will need to check for potential nonresponse bias resulting from the effects of these variables, especially if they will be part of an analysis.

## Acknowledgements

We thank Dr. Hebert Weisberg of the Political Science Department at The Ohio State University for commenting on an earlier draft of this paper. The survey upon which this paper is based was funded by the Ohio Agricultural Research and Development Center (OARDC), a research center within the College of Food, Agriculture, and Environmental Sciences at The Ohio State University. These acknowledgments notwithstanding, we are solely responsible for the content of this paper.

## References

- Audirac, I. (1999). Unsettled views about the fringe: Rural-urban or urban-rural frontiers. In *Contested Countryside: The Rural Urban Fringe in North America*, (Eds., O.J. Furuseth and M.B. Lapping). Brookfield, VT: Ashgate. 7-32.
- Brehm, J. (1993). *The Phantom Respondent*. Ann Arbor: University of Michigan Press.
- Cameron, A.C., and Trivedi, P.K. (2009). *Microeconometrics using Stata*. College Station, Texas: Stata Press.
- Church, A.H. (1993). Estimating the effects of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57, 62-79.



- Curtin, R., Presser, S. and Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413-28.
- Curtin, R., Presser, S. and Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Survey*, 69-1, 87-98.
- Davern, M., Rockwood, T.H., Sherrod, R. and Campbell, S. (2003). Prepaid monetary incentives and item nonresponse in face-to-face interviews. *Public Opinion Quarterly*, 67, 139-147.
- Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley & Sons, Inc.
- Dillman, D.A. (2000). *Mail and Internet Surveys: The Tailored Design Method*. New York: John Wiley & Sons, Inc.
- Dillman, D.A., Eltinge, J., Groves, R. and Little, R. (2002). Survey nonresponse in design, data collection, and analysis. In *Survey nonresponse*, (Eds., Groves *et al.*). New York: John Wiley & Sons, Inc.
- Dunlap, R.E., and Heffernan, R.B. (1975). Outdoor recreation and environmental concern: An empirical examination. *Rural Sociology*, 40, 18-30.
- Fowler, Jr. F.J. (2002). *Survey Research Methods*. Third Edition. Thousand Oaks, CA: Sage.
- Fox, R.J., Crask, M.R. and Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *The Public Opinion Quarterly*, 52-4, 467-491.
- Freudenburg, W.R. (1991). Rural-urban differences in environmental concern: A closer look. *Sociological Inquiry*, 61-2, 167-198.
- Fuchs, C., and Kenett, R. (2007). Missing data and imputation. In *Encyclopedia of Statistics in Quality and Reliability*. (Eds., F. Ruggeri, R.S. Kenett and F. Faltin). Wiley. 1090-1099.
- Gouldner, A. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25-2, 161-178.
- Goyder, J.C. (1982). Further evidence of factors affecting response rates to mailed questionnaires. *American Sociological Review*, 47, 550-553.
- Groves, R.M., and Cooper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.
- Groves, R.M., Singer, E. and Coming, A. (2000). Leverage-saliency theory of survey participation. *Public Opinion Quarterly*, 64, 299-308.
- Groves, R.M., Fowler, Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. Hoboken, NJ: John Wiley & Sons, Inc.
- Groves, R.M., Couper, M.P., Presser, S., Singer, E., Tourangeau, R., Acosta, G.P. and Nelson, L. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly*, 70-5, 139-147.
- Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70-5, 646-675.
- Hansen, R.A. (1980). A self-perception interpretation of the effects of monetary and nonmonetary incentives on mail survey response behavior. *Journal of Marketing Research*, 17, 77-83.
- James, J.M., and Bolstein, R. (1992). Large monetary incentives and their effect on mail survey response rates. *Public Opinion Quarterly*, 56, 442-453.
- Kaldenberg, D.O., Koenig, H.S. and Becker, B.W. (1994). Mail survey response patterns in a population of the elderly. *Public Opinion Quarterly*, 58-1, 68-76.
- Keeter, S., Miller, C., Kohut, A., Groves, R.M. and Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-48.
- Lobao, L.M. (1990). *Locality and Inequality: Farm and Industry Structure and Socioeconomic Conditions*. Albany, NY: The State University of New York Press.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press.
- Nathan, G. (2001). Telesurvey methodologies for household surveys: A review and some thoughts for the future. *Survey Methodology*, 27-1, 7-31.
- Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- Sharp, J.S., and Adua, L. (2009). The social basis of agro-environmental concern: Physical versus social proximity. *Rural Sociology*.
- Sharp, J.S., and Clark, J.K. (2008). Between the country and the concrete: Rediscovering the rural-urban fringe. *City & Community*. 7-1, 61-79.
- Shettle, C., and Mooney, G. (1999). Monetary incentives in U.S. government surveys. *Journal of Official Statistics*, 15-2, 231-250.
- Singer, E., Van Hoewyk, J. and Maher, M.P. (1998). Does the payment of incentives create expectation effects? *The Public Opinion Quarterly*, 62-2, 152-164.
- Singer, E., Van Hoewyk, J. and Maher, M.P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64, 171-188.
- Singer, E. (2006). Introduction: Nonresponse bias in household surveys. *Public Opinion Quarterly*, 70-5, 637-645.
- Teitler, J.O., Reichman, N.E. and Sprachman, S. (2003). Costs and benefits of improving response rate for a hard-to-reach population. *Public Opinion Quarterly*, 67, 126-138.
- Trussell, N., and Lavrakas, P.J. (2004). The influence of incremental increases in token cash incentives on mail survey response. *Public Opinion Quarterly*, 68-3, 349-367.
- Visser, P.S., Krosnick, J.S., Marquette, J. and Curtin, M. (1996). Mail surveys for election forecasting? An evaluation of the columbus dispatch poll. *Public Opinion Quarterly*, 60, 181-227.
- Weisberg, H.F. (2005). *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago and London: The University of Chicago Press.



- Williams, R. (2006). Generalized ordered logit/Partial proportional odds models for ordinal dependent variables. *The Stata Journal*, 6-1, 58-82.
- Yammarino, F.J., Skinner, S.J. and Childers, T.L. (1991). Understanding mail survey response behavior: A meta-analysis. *The Public Opinion Quarterly*, 55-4, 613-639.
- Yu, J., and Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research*, 20, 36-44.



# Evaluating within household selection rules under a multi-stage design

Tom Krenzke, Lin Li and Keith Rust<sup>1</sup>

## Abstract

The 2003 National Assessment of Adult Literacy (NAAL) and the international Adult Literacy and Lifeskills (ALL) surveys each involved stratified multi-stage area sample designs. During the last stage, a household roster was constructed, the eligibility status of each individual was determined, and the selection procedure was invoked to randomly select one or two eligible persons within the household. The objective of this paper is to evaluate the within-household selection rules under a multi-stage design while improving the procedure in future literacy surveys. The analysis is based on the current US household size distribution and intraclass correlation coefficients using the adult literacy data. In our evaluation, several feasible household selection rules are studied, considering effects from clustering, differential sampling rates, cost per interview, and household burden. In doing so, an evaluation of within-household sampling under a two-stage design is extended to a four-stage design and some generalizations are made to multi-stage samples with different cost ratios.

Key Words: Intraclass correlation; Design effects; Multi-stage sampling.

## 1. Introduction

The 2003 National Assessment of Adult Literacy (NAAL), conducted for the National Center for Education Statistics, provided an indicator of the nation's progress in English literacy for researchers, practitioners, policymakers, and the general public. As in the 1992 National Adult Literacy Study (NALS), adults were assessed in households in prose, document and quantitative literacy. The booklet designs were based on the 1992 NALS to allow for the measurement of trends between 1992 and 2003.

In order to reduce the cost of interviewers traveling to households, the NAAL involved a stratified four-stage cluster design that resulted in 18,500 completed assessments administered to adults age 16 and older. In the NAAL, counties were grouped to form Primary Sampling Units (PSUs), which were stratified and selected in the first stage. In the second stage, Secondary Sampling Units (SSUs) were formed and selected within the sampled PSUs. The SSUs were individual census blocks, or groups of adjacent blocks with at least 60 households (HHs) formed within tract boundaries. Subsequently, households were selected within SSUs, and one sample person (1 SP) was randomly selected for household sizes up to 3 ( $B \leq 3$ ), and two persons (2 SPs) were selected for household sizes greater than 3 ( $B > 3$ ), where  $B$  denotes the number of eligible persons per household. This rule followed the within-household sampling approach used in the first cycle of NAAL (NCES 2001), conducted in 1992. An evaluation of the selection rule was

conducted using the current US household size distribution and intraclass correlation coefficients computed from the 2003 survey. In doing so, an evaluation of within-household sampling under a two-stage design (Clark and Steel 2007) is extended to a four-stage design, as used in the NAAL survey and some generalizations are made to multi-stage samples with different cost ratios.

The data used for the evaluation include literacy measures from three scales derived from three types of literacy - prose, document, and quantitative. For more information about the NAAL types of literacy, refer to [http://nces.ed.gov/NAAL/fr\\_tasks.asp](http://nces.ed.gov/NAAL/fr_tasks.asp). Two types of estimates are used; averages (e.g., average prose literacy score) and percentage of adults at some level of literacy (e.g., percentage *Below Basic* prose literacy). For a discussion of the literacy levels used in NAAL, see [http://nces.ed.gov/NAAL/perf\\_levels.asp](http://nces.ed.gov/NAAL/perf_levels.asp). In addition to the NAAL data, the evaluation also uses US sample data from the international Adult Literacy and Lifeskills (ALL), which was conducted by Statistics Canada. The US sample in 2003, sponsored by NCES, was part of a comparative study that measured the skills of adults in several countries. Similar to the NAAL, the ALL was a multi-stage clustered sample survey and measured prose and document literacy, as well as numeracy (OECD 2005). The NAAL sample was much larger (18,500 completes) than the ALL sample (3,400 completes), and the target population for NAAL included ages 16+ while the target population for ALL included 16 to 65 year olds. Table 1 provides a summary of each survey's design and structure.

1. Tom Krenzke, Statistical Group, Westat, Rockville, Maryland 20850. E-mail: [tomkrenzke@westat.com](mailto:tomkrenzke@westat.com); Lin Li, Statistical Group, Westat, Rockville, Maryland 20850. E-mail: [linli@westat.com](mailto:linli@westat.com); Keith Rust, Statistical Group, Westat, Rockville, Maryland 20850. E-mail: [keithrust@westat.com](mailto:keithrust@westat.com).



**Table 1**  
**Features of the NAAL and ALL surveys**

Survey	Area sample	Completes	Data collection	Assessments	Ages	Within-HH sampling rule
NAAL	PSUs, SSUs households, Persons	18,500	Screener Interview Assessment	Prose Document Quantitative	16+	$B \leq 3, b = 1$ $B > 3,$ $b = 2$
ALL	PSUs, SSUs, households, Persons	3,400	Screener Interview Assessment	Prose Document Numeracy	16-65	$B \leq 3, b = 1$ $B > 3, b = 2$

Note: PSU = Primary Sampling Unit, SSU = Secondary Sampling Unit,  $b$  = sample size,  $B$  = household size.

A discussion of the design considerations that helped form the evaluation of the within-household sampling rules is provided in Section 2. Section 3 discusses the computation of intra-household correlations under multi-stage sample designs and focuses on incorporating the clustering impact from the initial stages of sample selection when deciding on a within-household selection rule. An evaluation of selection rules was conducted using data from the in-person adult literacy surveys and the results are provided in Section 4. Finally, a brief summary is given in Section 5.

## 2. Design considerations

There are a number of factors that need to be considered when evaluating the within-households selection rules for surveys such as NAAL and ALL. The remainder of this section will discuss the impact of the following factors on within-household sampling: household burden, clustering persons within households, differential sampling rates, multi-stage sampling, cost considerations, computerized systems, domains of interest and household composition.

*Household burden.* For the adult literacy surveys, the interview and the assessment take about an hour and a half to administer in total. Therefore, one concern about selecting more than one person per household is the increase of burden to the household and the impact on response rates. However, there is no significant difference (0.05 significance level) in the refusal rates between 1- and 2-SP households in ALL and NAAL as shown in Table 2.

*Clustering persons within households.* Kish (1965) discusses the benefits of a cluster sample to a simple random sample. A cluster sample typically has a lower cost per person, however the unit variance is higher and it causes greater complexities in statistical analysis. Kish introduced the concept of a design effect (DEFF), which measures the increase in variance due to deviations from a simple random sample, such as clustering persons within households. Many surveys limit the selection to one sample person (SP) per household because of concerns over the increased clustering effect (*i.e.*, increasing effect on variance estimates) associated with multiple SPs per household. The DEFF due to

clustering can be expressed as:  $DEFF_{clu} = 1 + (\bar{b} - 1) \text{Rho}$ , where  $\bar{b} = \sum (M_B / M) b_B$ ,  $M_B$  = number of households of size  $B$ ,  $M$  = number of households, and  $b_B$  = sample size of persons within households of size  $B$  (Kish 1965). This DEFF component increases when the sample size within a household increases or when the value of the intraclass correlation (Rho) increases. As given in Cochran (1977), Rho can be approximated as:

$$\text{Rho} = 1 - \frac{\sigma_w^2}{\sigma^2},$$

where

$$\sigma_w^2 = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_i)^2 / (n - a),$$

and

$$\sigma^2 = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y})^2 / (n - 1),$$

where  $a$  is the number of sampled households, and  $b$  is the number of sampled persons per household. The DEFF due to clustering is examined further for different within-household sampling rules in the next section.

*Differential sampling rates.* A clustering effect is not the only factor that increases the variance. Increases in variance are also due to differential sampling rates (resulting in differential weights). Under a 1 SP per household strategy, the increase is directly related to the variation in household size since the sampling rate could vary from 1 out of 1 to 1 out of 7 or more. The DEFF due to differential sampling rates is expressed as:  $DEFF_{wgt} = \sum (p_B / k_B) \sum (p_B k_B)$ , where  $p_B = N_B / N$ ,  $N_B$  = number of eligible persons in the population in households of size  $B$ ,  $N$  = number of eligible persons in the population, and  $k_B$  = sampling rate within households of size  $B$  (Kish 1965). Under certain conditions, the overall DEFF can be expressed as the product of the clustering and differential sampling rate components:  $DEFF = DEFF_{clu} \times DEFF_{wgt}$ . Kalton, Brick and Lê (2005) suggest this product is applicable when the weights are random or approximately random.

Table 2  
Refusal rates by 1- and 2-SP households for the adult literacy surveys

Survey	Subgroup	Refusal rate %
NAAL	1-SP households	16.3
	2-SP households	15.7
ALL	1-SP households	17.6
	2-SP households	16.2

Note: SP = sample person.

To arrive at a self-weighting sample, persons within households would need to be selected at a constant rate. However, a rate-based approach is not preferred in most surveys since it would result in walking away from a portion of single-person households and, thus, would increase the cost of the survey. We limit the alternative rules under consideration to those with a minimum of 1 SP per household. Out of concern for burdening households, the maximum sample size was set to two. The sampling rules under consideration are:

- 1. Take1: 1 SP no matter the household size.
- 2. Rule2: 1 SP for household sizes up to 2; otherwise 2 SPs are selected.
- 3. NAAL3: 1SP for household sizes up to 3; otherwise 2 SPs are selected.
- 4. Rule4: 1 SP for household sizes up to 4; otherwise 2 SPs are selected.
- 5. Frac5: take at least 1 SP, but no more than 2 SPs and the sample size is a fraction. That is, if the sample size for a household with two eligible persons is 1.6, then two persons are selected 60 percent of the time at random, and one person is selected 40 percent of the time.

While the Take1 approach does not attempt to reduce the DEFF due to differential sampling rates, it is not subject to a clustering impact. However, the other four approaches listed above provide a reduction in the differential sampling rate component while introducing a clustering effect. In the case of Frac5, under the assumption that  $\pi$ -weights are used, as assumed throughout this paper, the approach would result in the most reduction in the differential sampling rate component. The  $\pi$ -weights approach is based on the unconditional selection probability of the person within the household. If the actual sample size within a household is used in the form of ratio weights, the differential sampling rate increases the benefit is less clear and depends on Rho. Figure 1 illustrates the best options under a two-stage household design with fixed effective sample size of persons, without any cost considerations. The US national household size distribution from the 2007 Current Population Survey was used for this illustration. As shown in Figure 1, the fractional approach is the best rule for a wide range of values of Rho. The fractional approach can be programmed into a computerized system when enumerating and selecting household members (more discussion on computerized

systems follows). If computerized systems are not available for screening, then the best approach for low values of Rho is the more clustered approach, Rule2; and the NAAL3 rule is best for Rho values greater than about 0.34.

*Multi-stage sampling.* For multi-stage area designs, the clustering impact of sampling within households is affected by the clustering due to PSUs and SSUs. As pointed out by Kish (1965), the clustering of households and persons within PSUs and SSUs increases the sampling variance (*i.e.*, units within PSUs and SSUs are more similar to each other). The incremental impact of clustering within households may be dampened by the domination of the PSU and SSU variance components (however, the magnitude of the impact will differ depending on the type of estimate and variable). That is, more persons within a household can be selected for surveys with a large amount of clustering due to the first two stages of sampling. Details of this distinction are provided in Section 3.

*Cost considerations.* The cost of screening a household in a 1 SP per household design versus the cost of interviewing/assessing a second person in a household is investigated in an extensive analysis presented later .

*Computerized systems.* Computerized systems, such as Computer-Assisted Personal Interview (CAPI), have the capability of handling fractional sample sizes. That is, the random selection of 1 or 2 SPs given a pre-assigned fractional sample size can be programmed. Computerized systems also have the capability of sorting the list of eligible persons and selecting 2 SPs with a systematic random sample. Another benefit is that the selection program can be tested and validated prior to data collection.

*Domains of interest.* As mentioned earlier, optimal within household sampling depends on the magnitude of the clustering effect associated with the variable of interest. The clustering effect may be much smaller when the variable is associated with a subgroup of the population, rather than the entire population. For example, when a key reporting domain is gender in a survey of the adult population, the reporting category of males is likely to have an average of 1 SP per household and less likely to have 2 male SPs which would introduce a clustering effect. Therefore, when there are multiple domains of interest in a typical household, it is often beneficial to select more than 1 SP within a household. Refer to Mohadjer and Curtin (2008) for an example of design considerations for a survey with focus on multiple subgroups of the population.

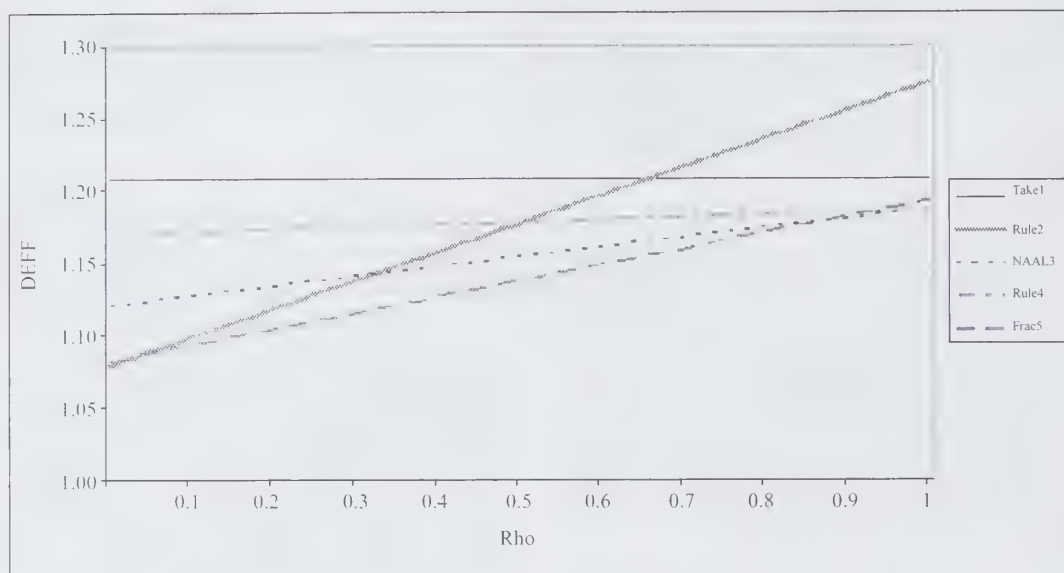


Figure 1 Initial analysis of within-household selection rules

*Household composition.* Lastly, one may want to consider the household composition and relationships of persons within a household when devising the selection rule. Table 3 displays values of Rho for various relationships between household members, for household with 2 SPs in the NAAL survey. Rho varies greatly by household member relationships. The relationships were derived from gender and age.

### 3. Estimation of intra-household Rho and DEFF under multi-stage sampling

The discussion about Rho thus far has been related to a two-stage design, but both NAAL and ALL have four stages of sampling. The total variance can be decomposed into four between-variance terms attributable to PSUs, SSUs, households and persons, as follows:

$$\sigma_T^2 = \sigma_{\text{PSU}}^2 + \sigma_{\text{SSU(PSU)}}^2 + \sigma_{\text{HH(SSU)}}^2 + \sigma_{\text{PERS(HH)}}^2.$$

As shown below, when applying a two-stage approach to estimate Rho for a four-stage sample design, the numerator not only contains the between household component, but also contains contributions from the between PSU and between SSU components inflating the values of Rho for our purpose.

$$\text{Rho} = 1 - \frac{\sigma_{\text{PERS(HH)}}^2}{\sigma_T^2} = \frac{\sigma_{\text{PSU}}^2 + \sigma_{\text{SSU(PSU)}}^2 + \sigma_{\text{HH(SSU)}}^2}{\sigma_T^2}.$$

Therefore, when evaluating rules for within-household sampling under a multi-stage design, we assume the PSU and SSU design will be the same in the future. This can be accomplished by limiting our focus to within SSU sampling. Therefore, the computation of Rho is contained within SSUs, that is, it is done in a compact manner without effect from the PSU and SSU components. We refer to this as the compact (*i.e.*, within SSU) Rho denoted by  $\text{Rho}^*$ , expressed as:

$$\text{Rho}^* = \frac{\sigma_{\text{HH(SSU)}}^2}{\sigma_{\text{HH(SSU)}}^2 + \sigma_{\text{PERS(HH)}}^2}.$$

Using the compact  $\text{Rho}^*$ , we now derive the estimated DEFF under a multi-stage sample design for the purpose of determining optimal within-household sample sizes. The variance of an estimate ( $\hat{\theta}$ ) with  $b$  persons per household can be decomposed as:

$$\text{Var}(\hat{\theta}) = \frac{\sigma_{\text{PSU}}^2}{n_{\text{PSU}}} + \frac{\sigma_{\text{SSU(PSU)}}^2}{n_{\text{SSU}}} + \frac{\sigma_{\text{HH(SSU)}}^2}{n_{\text{HH}}} + \frac{\sigma_{\text{PERS(HH)}}^2}{bn_{\text{HH}}}$$

where,  $n_{\text{PSU}}$ ,  $n_{\text{SSU}}$ ,  $n_{\text{HH}}$  and  $bn_{\text{HH}}$  are the sample sizes of PSUs, SSUs, households and persons, respectively.



**Table 3**  
**Rho for NAAL assessment scores by household member relationships**

Estimate	Siblings	Child-guardian	Married	Others
Number of households with 2 SPs	111	205	180	434
Average prose score	0.42	0.35	0.70	0.59
Average document score	0.40	0.27	0.72	0.54
Average quantitative score	0.46	0.36	0.63	0.56
Percentage Below Basic prose	0.52	0.41	0.79	0.67
Percentage Below Basic document	0.54	0.40	0.78	0.60
Percentage Below Basic quantitative	0.51	0.41	0.77	0.65

Then the DEFF due to clustering, relative to taking one person per household and  $bn_{HH}$  households is:

$$\begin{aligned}
 DEFF_{clu}^{HH} &= \frac{\frac{\sigma_{PSU}^2}{n_{PSU}} + \frac{\sigma_{SSU(PSU)}^2}{n_{SSU}} + \frac{\sigma_{HH(SSU)}^2}{n_{HH}} + \frac{\sigma_{PERS(HH)}^2}{bn_{HH}}}{\frac{\sigma_{PSU}^2}{n_{PSU}} + \frac{\sigma_{SSU(PSU)}^2}{n_{SSU}} + \frac{\sigma_{HH(SSU)}^2}{bn_{HH}} + \frac{\sigma_{PERS(HH)}^2}{bn_{HH}}} \\
 &= \frac{\frac{\sigma_{PSU}^2}{n_{PSU}} + \frac{\sigma_{SSU(PSU)}^2}{n_{SSU}} + \frac{1}{bn_{HH}}(\sigma_{HH(SSU)}^2 + \sigma_{PERS(HH)}^2) + (b-1)\sigma_{HH(SSU)}^2)}{\frac{\sigma_{PSU}^2}{n_{PSU}} + \frac{\sigma_{SSU(PSU)}^2}{n_{SSU}} + \frac{1}{bn_{HH}}(\sigma_{HH(SSU)}^2 + \sigma_{PERS(HH)}^2)} \\
 &= \frac{bn_{HH} \left( \frac{\sigma_{PSU}^2}{n_{PSU}} + \frac{\sigma_{SSU(PSU)}^2}{n_{SSU}} \right) + (1 + (b-1) Rho^*)}{\frac{\sigma_{HH(SSU)}^2 + \sigma_{PERS(HH)}^2}{bn_{HH} \left( \frac{\sigma_{PSU}^2}{n_{PSU}} + \frac{\sigma_{SSU(PSU)}^2}{n_{SSU}} \right)} + 1} \\
 &= \frac{k^* + (1 + (b-1) Rho^*)}{k^* + 1}
 \end{aligned}$$

where,

$$\begin{aligned}
 k^* &= \frac{bn_{HH} \left( \frac{\sigma_{PSU}^2}{n_{PSU}} + \frac{\sigma_{SSU(PSU)}^2}{n_{SSU}} \right)}{(\sigma_{HH(SSU)}^2 + \sigma_{PERS(HH)}^2)} \\
 &= \frac{\frac{\sigma_{PSU}^2}{n_{PSU}} + \frac{\sigma_{SSU(PSU)}^2}{n_{SSU}}}{\frac{1}{bn_{HH}}(\sigma_{HH(SSU)}^2 + \sigma_{PERS(HH)}^2)}
 \end{aligned}$$

Alternatively,  $DEFF_{clu}^{HH}$  can be expressed as:

$$\begin{aligned}
 DEFF_{clu}^{HH} &= 1 + \frac{(b-1) Rho^*}{k^* + 1} \\
 &= 1 + (b-1) Rho^{**}
 \end{aligned}$$

where,

$$\begin{aligned}
 Rho^{**} &= \frac{Rho^*}{k^* + 1} \\
 &= \frac{\left( \frac{\sigma_{HH(SSU)}^2}{\sigma_{HH(SSU)}^2 + \sigma_{PERS(HH)}^2} \right) \frac{1}{bn_{HH}} (\sigma_{HH(SSU)}^2 + \sigma_{PERS(HH)}^2)}{\frac{\sigma_{PSU}^2}{n_{PSU}} + \frac{\sigma_{SSU(PSU)}^2}{n_{SSU}} + \frac{\sigma_{HH(SSU)}^2}{bn_{HH}} + \frac{\sigma_{PERS(HH)}^2}{bn_{HH}}} \\
 &= \frac{\frac{\sigma_{HH(SSU)}^2}{bn_{HH}}}{\frac{\sigma_{PSU}^2}{n_{PSU}} + \frac{\sigma_{SSU(PSU)}^2}{n_{SSU}} + \frac{\sigma_{HH(SSU)}^2}{bn_{HH}} + \frac{\sigma_{PERS(HH)}^2}{bn_{HH}}} \\
 &= \frac{\sigma_{HH(SSU)}^2}{\frac{bn_{HH} \sigma_{PSU}^2}{n_{PSU}} + \frac{bn_{HH} \sigma_{SSU(PSU)}^2}{n_{SSU}} + \sigma_{HH(SSU)}^2 + \sigma_{PERS(HH)}^2}
 \end{aligned}$$

The  $Rho^{**}$  measure is a useful expression for the intra-household correlation under a multi-stage design, which is equal to  $Rho^*$  when  $\sigma_{PSU}^2 = \sigma_{SSU(PSU)}^2 = 0$ . The compact  $Rho^*$  measure is useful for evaluating optimal sample sizes while varying the variance ratio  $k^*$ . Note, however, that in general  $Rho^{**}$  is a function of  $n_{PSU}$ ,  $n_{SSU}$  and the total sample size of persons, whereas  $Rho^*$  does not depend on these.

As shown in Table 4, the variance ratio  $k^*$ , which is the variance from the first two stages divided by the variance from the last two stages, for a one person per household design, ranges from 0.68 to 1.61 across types of assessments and estimates for the ALL survey.

Table 5 shows estimates for  $Rho$  (computed under a two-stage design assumption), the compact  $Rho^*$  and  $Rho^{**}$  (computed under a multi-stage design assumption where  $k^* = 1$ ) for average NAAL and ALL literacy assessment scores. When including the clustering impact from the first two stages of the four-stage design, the values of the compact  $Rho^*$  and  $Rho^{**}$  are much smaller than  $Rho$ . For example, the two-stage  $Rho$  for the NAAL average prose score is 0.57 and the compact  $Rho^*$  is equal to 0.33 and  $Rho^{**}$  is equal to 0.17. The table also shows that values of the compact  $Rho^*$  for average scores are at about the same level for NAAL (range from 0.32 to 0.33) and ALL (range

from 0.29 to 0.39). There is some variation by the type of estimate as well; values of  $Rho^*$  for ALL are 0 to 0.2 lower for the percentage in Level 1 or 2 than for the average scores. Values of  $Rho^*$  can also vary by household size as shown in Figure 2 in Appendix A.

#### 4. Evaluation and results

We compared the current sampling rules with optimal sampling rules by minimizing a variance-cost (VC) function, which is the product of the DEFFs (*i.e.*, variance increase) due to clustering and weighting, and a cost function that is used by Kish (1965):

$$VC = DEFF_{clu}^{HH*} \times DEFF_{wgt} \times n \left( c_p + \frac{c_{HH}}{b} \right),$$

where  $c_p$  = cost per added person and  $c_{HH}$  = cost per added household. Note that  $n/\bar{b}$  represents the number of sampled households. To account for the differential clustering effects for each household size  $B$ , we replace  $DEFF_{clu}^{HH}$  with:

$$DEFF_{clu}^{HH*} = \frac{k^* + \sum_B \frac{M_B}{M} (1 + (b_B - 1) Rho_B^*)}{k^* + 1}$$

where  $Rho_B^*$  is computed as described in Appendix A.

Note that the VC function represents the additional cost of increasing the overall sample size to offset the increase in variance due to the DEFF components. Table 6 provides the results for optimal integer solutions as computed by a computational algorithm which is described in Appendix B. The table shows that as the cost ratio increases from 0.5 to 1 for  $k^* = 1$ , we would want to take more persons per household, that is, 2 out of 2 instead of 1 out of 2. As the variance ratio goes from 1 to 3 for optimal integer solutions,

the only change is for household size of 2 and cost ratio of 0.5. That is, when the variance ratio is equal to 3, it is beneficial to take 2 out of 2 instead of 1 out of 2.

Table 6 also gives the results when fractional sample sizes are allowed. The variance and cost ratios for NAAL and ALL tend to be about 1, where it appears that selecting 1 out of 1, 1.6 out of 2, and 2 otherwise is the best rule. The effects of cost and variance ratios are clearer under the fractional sample sizes when compared to the integer solutions.

If the cost of conducting a screener is small in relation to the cost of interviewing, then variances can be reduced using the fractional walk-away approach. Table 6 shows optimal walk-away sample sizes. Under this approach, for example, a sample size of 0.9 indicates that we walk away from 10 percent of the households where  $B = 1$ . If the cost of screening is a very small portion of the cost of interviewing, then the optimal design may involve walking away from many more households.

Under the likely NAAL/ALL parameters for cost ratios ( $C_{HH}/C_p = 1$ ) and variance ratios ( $k^* = 1$ ), when compared to the Take1 approach, the VC function can be reduced by about 9 percent by using the NAAL/ALL sampling rule, 19 percent by using the optimal integer solution, 20.4 percent using the optimal fractional solution, and 20.6 using the optimal walk-away approach. In general, the gains from deviating from the Take1 approach grow as the cost per additional households (*i.e.*, screening) increases. The average cluster sizes for each approach are given in Table 7. For the NAAL and optimal integer rule, the average cluster size indicates the percentage of households with 2 SPs. For example about 6 percent of the households would have 2 SPs under the NAAL3 strategy.

**Table 4**  
Values of  $k^*$  for the ALL sample

ALL estimate	$k^*$
Average prose score	0.95
Average document score	1.56
Average quantitative/numeracy score	1.13
Percentage in Level 1 or 2 prose	0.68
Percentage in Level 1 or 2 document	1.61
Percentage in Level 1 or 2 numeracy	1.10

**Table 5**  
Values for  $Rho$ ,  $Rho^*$ , and  $Rho^{**}$  for literacy assessment scores

Estimate	Rho		Rho*		Rho**	
	NAAL	ALL	NAAL	ALL	NAAL	ALL
Number of households with 2 SPs	930	162	930	162	930	162
Average prose score	0.57	0.60	0.33	0.38	0.17	0.19
Average document score	0.53	0.50	0.33	0.29	0.17	0.15
Average quantitative/numeracy score	0.54	0.58	0.32	0.39	0.16	0.20
Percentage Below Basic(NAAL)/Level 1 or 2 (ALL) prose	0.65	0.44	0.42	0.28	0.21	0.14
Percentage Below Basic (NAAL)/Level 1 or 2 (ALL) document	0.61	0.37	0.39	0.28	0.20	0.14
Percentage Below Basic quantitative (NAAL)/Level 1 or 2 (ALL) numeracy	0.62	0.36	0.40	0.17	0.20	0.09

Note:  $Rho^{**}$  is computed assuming  $k^* = 1$ .

**Table 6**  
Optimal expected number of persons per household by type of person sampling method and household size (B)

$k^*$	$C_{HH} / C_p$	Person Sampling Method											
		Integer				Fractional				Walk-away			
		$B = 1$	$B = 2$	$B = 3$	$B = 4$	$B = 1$	$B = 2$	$B = 3$	$B = 4$	$B = 1$	$B = 2$	$B = 3$	$B = 4$
1	0.5	1	1	2	2	1	1.4	2	2	0.6	1.3	2	2
1	1	1	2	2	2	1	1.6	2	2	0.9	1.6	2	2
1	2	1	2	2	2	1	1.9	2	2	1	1.9	2	2
3	0.5	1	2	2	2	1	1.6	2	2	0.8	1.5	2	2
3	1	1	2	2	2	1	1.8	2	2	1	1.8	2	2
3	2	1	2	2	2	1	2	2	2	1	2	2	2

**Table 7**  
Percent reduction of NAAL3 and optimal solutions from Take1 strategy and average cluster sizes

$k^*$	$C_{HH} / C_p$	Percentage reduction from Take1 strategy				Average cluster sizes			
		NAAL3	Integer	Fractional	Walk-away	NAAL3	Integer	Fractional	Walk-away
1	0.5	8.2	13.0	15.8	18.0	1.06	1.18	1.38	1.21
1	1	9.1	19.2	20.4	20.6	1.06	1.68	1.48	1.45
1	2	9.9	26.1	26.1	26.1	1.06	1.68	1.63	1.63
3	0.5	8.6	17.3	18.7	19.0	1.06	1.68	1.48	1.37
3	1	9.5	23.7	23.9	23.9	1.06	1.68	1.58	1.58
3	2	10.4	30.2	30.2	30.2	1.06	1.68	1.68	1.68

Lastly, a sensitivity analysis was conducted by varying the values of  $Rho^*$ . A regression model was fit on the percentage reduction from the Take1 strategy of the VC function, with the independent variables being the approach (NAAL3, integer, fractional, walk-away), cost ratio (0.1, 0.5, 1, 2, 10), variance ratio (1, 3, 5) and  $Rho^*$  (+/- 0.1). For the range of data,  $Rho^*$  had a limited impact (parameter estimate -7.4 with an associated standard error of 4.5) on the percentage reduction of the VC function, while the other factors had more of an impact.

## 5. Summary

Several design considerations were taken into account when evaluating the within - household selection rule for the NAAL and ALL surveys, including taking into account clustering effects from initial stages of sampling. To facilitate the evaluation, we formulate a way to incorporate PSU and SSU variance contributions into the computation of the DEFF due to clustering and the intra-household correlation when deciding how many persons and how many households to select in a multi-stage sample design. In doing so, we introduce compact  $Rho^*$  measure, which is computed within the SSU so it is not impacted by the PSU and SSU variance components. This is useful when determining the DEFF due to clustering within households, while varying the contribution to the total variance from the PSU and SSU stages of selection in multi-stage sample

designs. The measure  $Rho^{**}$  is introduced as an expression for the intra-household correlation under a multi-stage design, taking into consideration the contribution to total variance from the first two stages of selection.

In addition, a computational algorithm was developed to compute optimal sample size solutions, incorporating the DEFFs due to clustering, differential sampling rates, and costs.

In general, the main factors on the percentage reduction of the VC function from the Take1 approach are the level of dominance from the PSU and SSU variance components in multi-stage sampling, the cost ratio and the rule used. For the range of data evaluated,  $Rho^*$  had limited impact on the reduction in VC from the Take1 approach. In general, the NAAL rule improves on the widely-used Take1 approach. The optimal integer rule improves on the NAAL rule. However, the optimal fractional rule has limited gains over the optimal integer rule. The optimal walk-away rule has gains over the other rules for lower cost ratios. Lastly, when the first two variance components dominate and cost ratio is high, then the integer, fractional and walk-away rules are essentially the same.

## Acknowledgements

The authors acknowledge valuable contributions by Leyla Mohadjer and Bob Fay.



## Appendix A

### Estimates of $\text{Rho}^*$ by household size

Survey estimates are not attainable for  $\text{Rho}^*$  by household size since only 1 SP was selected for household size of 3 or less and since the sample size was too small to create estimates for each household size of 4 or more. Therefore, estimates of  $\text{Rho}^*$  by household size are modeled using Census data. Figure 2 shows  $\text{Rho}^*$  on the y-axis and household size on the x-axis. The upper line is from the US Census public-use microdata sample (PUMS) file for education attainment for ages 25+. The upper line shows that education attainment is more similar among households with two adults, perhaps more likely to be married couples. It shows a drop off when going from two to three adults. We captured the variation in households size by computing the ratio of  $\text{Rho}^*$  for the NAAL prose literacy scores to the  $\text{Rho}$  for the Census PUMS education attainment among households with  $B > 3$  and applying the ratio to the PUMS  $\text{Rho}$  across all household sizes. The resulting values are the estimates of compact  $\text{Rho}_B^*$  for  $B = 1, 2, \dots, 11$ .

## Appendix B

### Computational algorithm

A computational algorithm was developed to arrive at optimal within-household sample sizes for each household size  $B$ . The algorithm was constructed to generate optimal integer or fractional solutions that capture the effects of clustering, differential sampling rates and cost, under the constraints of at least one selected person per household and no more than 2. Here are the steps of the algorithm (all processing runs converged within four iterations):

- Initialize by setting  $b = 1$  for all values of  $B$  (Take1).
- Compute  $\text{DEFF}_{\text{clu}}^{\text{HH}^*}$ ,  $\text{DEFF}_{\text{wgt}}$ ,  $c_p$ ,  $c_{\text{HH}}$ , and  $\text{VC}(0)$ .
- Do  $I = 1$  to 5.
  - Do  $B = 1$  to 11.
    - Compute  $\text{DEFF}_{\text{clu}}^{\text{HH}^*}$ ,  $\text{DEFF}_{\text{wgt}}$ ,  $c_p$ ,  $c_{\text{HH}}$ , and  $\text{VC}$  for all  $1 \leq b_B \leq 2$ , given the set of  $b_B$ , for all  $B' \neq B$ .
    - Identify the  $b_B$  with the smallest value of  $\text{VC}$ .
  - End.
  - If  $\text{VC}(I) = \text{VC}(I - 1)$  then stop.
- End.

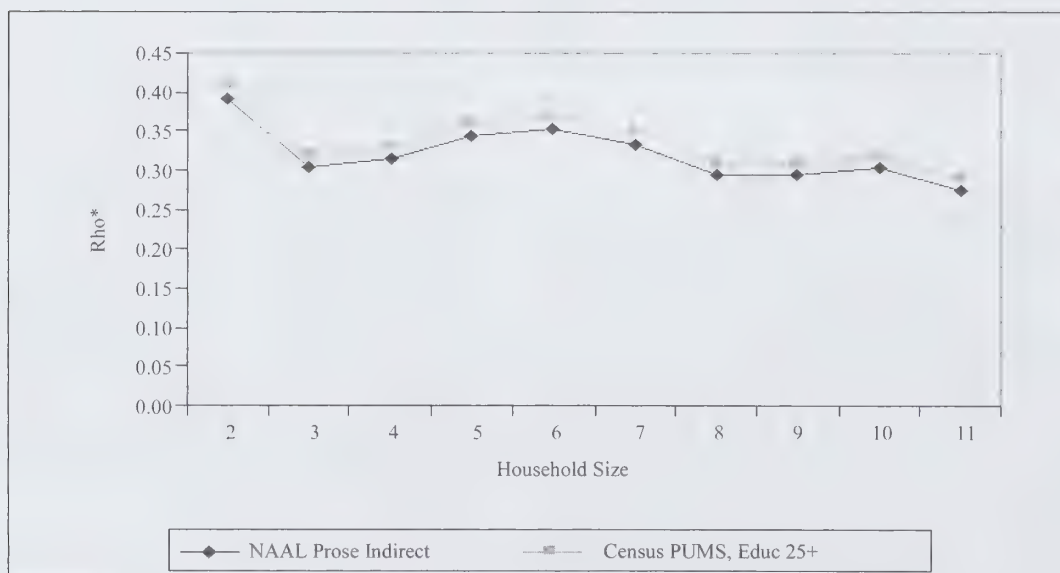


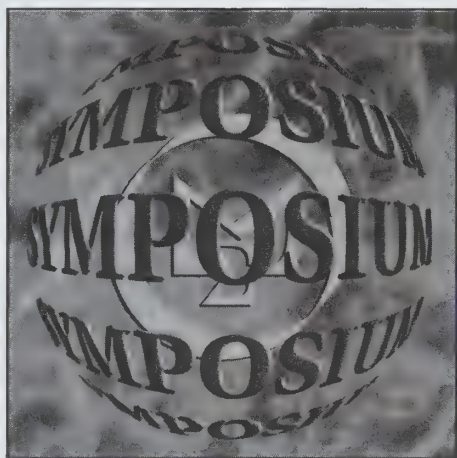
Figure 2 Estimates of  $\text{Rho}^*$  for NAAL by household size

## References

- Clark, R.G., and Steel, D.G. (2007). Sampling within Households in Household Surveys. *Journal of the Royal Statistical Society, Series A*, 170, 63-82.
- Cochran, W.G. (1977). *Sampling Techniques*. 3<sup>rd</sup> Ed. New York: John Wiley & Sons, Inc.
- Kalton, G., Brick, J.M. and Lê, T. (2005). Estimating Components of Design Effects for Use in Sample Design, Household Sample Surveys in Developing and Transition Countries, Chapter VI United Nations, New York, 95-121.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Mohadjer, L., and Curtin, L.R. (2008). Balancing sample design goals for the National Health and Nutrition Examination Survey. *Survey Methodology*, 34, 1, 119-126.
- NCES (2001). Technical Report and Data File User's Manual For the 1992 National Adult Literacy Survey. U.S. Department of Education, National Center for Education Statistics, Washington, DC.
- OECD (2005). Learning a Living: First Results of the Adult Literacy and Life Skills Survey. Organisation for Economic Co-operation and Development, Paris. Statistics Canada, Ottawa.







## 2010 International Methodology Symposium

Statistics Canada  
October 26-29, 2010  
Ottawa, ON, Canada

### **Social Statistics: The Interplay among Censuses, Surveys and Administrative Data**

Statistics Canada's 2010 International Methodology Symposium will take place at the Crowne Plaza Hotel, located in the heart of downtown Ottawa, from October 26-29, 2010.

The Symposium will be titled "**Social Statistics: The Interplay among Censuses, Surveys and Administrative Data**". Members of the statistical community, such as those from private organizations, governments, or universities, are invited to attend, particularly if they have a special interest in statistical or methodological issues resulting from the use of multiple sources of data (censuses, sample surveys or administrative data).

The first day will consist of workshops, while the following days will consist of both plenary and parallel sessions covering a variety of topics. Additional research and results may be presented via poster sessions.

The presentations will be related to the methodological aspects of using multiple sources of data. Topics may include:

- Sampling Frames and Sample Design
- Coordinating Samples
- Content and Questionnaire Design
- Data Collection Methods and Acquisition of Administrative Data
- Supplementing Survey Data with Administrative Data
- Administrative Data for Direct Estimation
- Statistical Databases from Administrative Data (e.g., Population Registers)
- Imputation
- Weighting and Estimation
- Dissemination and Data Access
- Record Linkage Techniques
- Record Linkage Software
- Measurement Errors
- Response Burden
- Treatment of Nonresponse
- Confidentiality, Privacy and Ethical Issues
- Small Area Estimation

Visit our Internet site regularly to obtain further details about the program, workshops, registration, accommodation, tourism information and more at

<http://www.statcan.gc.ca/conferences/symposium2010/index-eng.htm>



Statistics  
Canada

Statistique  
Canada

Canada

# ICES IV

## FOURTH INTERNATIONAL CONFERENCE ON ESTABLISHMENT SURVEYS (ICES IV) PLANNED FOR 2012

Planning is underway for the Fourth International Conference on Establishment Surveys (ICES IV). If you've attended any of the past conferences, you know how invaluable they have been to the literature and practice of establishment surveys. If you are newer to the establishment survey field, you will find the conference especially rewarding. Since the last ICES held in 2007, many new techniques have been developed by practitioners around the world. A major strength of the conferences is the strong international presence, both in the program development and attendance. Over 400 people from 94 countries attended ICES III. On June 11-14 2012, survey practitioners from government agencies, academia, private sector and more will gather at the Sheraton Centre Montreal in Quebec, Canada for ICES IV and continue the tradition of sharing innovative techniques and best practices to address common issues.

Sponsorship of the meetings is being provided by the American Statistical Association, ASA Section on Survey Research Methods, ASA Section on Government Statistics, International Association of Survey Statisticians, and the Statistical Society of Canada. Administrative support for ICES IV will be provided by the American Statistical Association, similar to previous ICES meetings. Also, many other organizations and government agencies are or will be providing support for the conference.

With the support of these many great organizations and the diverse gathering of individuals involved in establishment surveys, we anticipate that ICES IV will prove to be another fruitful conference in the valuable ICES series. So, save the date, **June 11-14, 2012**, and join practitioners from around the globe in **Montreal, Canada!** You can participate in the growing ICES IV program discussing current issues, future vision, and cutting-edge methods in surveying businesses, farms and institutions. Expect updates on participation and program details to ICES IV through this newsletter and the upcoming ICES IV website. Inquiries may be directed to [ices4@amstat.org](mailto:ices4@amstat.org).

# **JOURNAL OF OFFICIAL STATISTICS**

**An International Review Published by Statistics Sweden**

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## **Contents**

**Volume 25, No. 4, 2009**

Innovative Statistics to Improve Our Notion of Reality Henk K. van Tuinen .....	431
Research and Development in Official Statistics and Scientific Co-operation with Universities: A Follow-Up Study Risto Lehtonen, Carl-Erik Särndal .....	467
Does the Effect of Incentive Payments on Survey Response Rates Differ by Income Support History? Juan D. Barón, Robert V. Breunig, Deborah Cobb-Clark, Tue Gørgens, Anastasia Sartbayeva .....	483
Design of Web Questionnaires: The Effect of Layout in Rating Scales Vera Toepoel, Marcel Das, Arthur van Soest.....	509
The Effect of Single-Axis Sorting on the Estimation of a Linear Regression Matthias Schmid.....	529
Using Bayesian Networks to Create Synthetic Data Jim Young, Patrick Graham, Richard Penny .....	549
Evaluating Alternative One-Sided Coverage Intervals for a Proportion Yan K. Liu, Phillip S. Kott .....	569
Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB Establishment Survey Jörg Drechsler, J.P. Reiter.....	589
Editorial Collaborators .....	611

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)



## Volume 37, No. 4, December/décembre 2009

Juli ATHERTON, Benoit CHARBONNEAU, David B. WOLFSON, Lawrence JOSEPH, Xiaojie ZHOU, Alain C. VANDAL Bayesian optimal design for changepoint problems .....	495
Jingjing WU, Rohana J. KARUNAMUNI On minimum Hellinger distance estimation .....	514
Noomen Ben GHORBAL, Christian GENEST, Johanna NEŠLEHOVÁ On the Ghoudi, Khoudraji, and Rivest test for extreme-value dependence .....	534
Christopher R. BILDER, Thomas M. LOUGHIN Modeling multiple-response categorical data from complex surveys .....	553
Lajos HORVÁTH, Piotr KOKOSZKA, Matthew REIMHERR Two sample inference in functional linear models .....	571
Jianguo SUN, Junshan SHEN Efficient estimation for the proportional hazards model with competing risks and current status data .....	592
Haonan WANG, Jun ZHU Variable selection in spatial regression via penalized least squares .....	607
Rui WANG, Stephen W. LAGAKOS Inference after variable selection using restricted permutation methods .....	625
Liqun XI, Ray WATSON, Ji-Ping WANG, Paul S.F. YIP Estimation in capture-recapture models when covariates are subject to measurement errors and missing data .....	645
Guosheng YIN, Hui LI Least squares estimation of varying-coefficient hazard regression with application to breast cancer dose-intensity data .....	659

## Volume 38, No. 1, March/mars 2010

Paul GUSTAFSON	
Report from the previous editor.....	1
Jiahua CHEN	
Notes from the New Editor.....	5
Hung HUNG, Chin-Tsang CHIANG	
Estimation methods for time-dependent AUC models with survival data.....	8
Fang YAO, Radu V. CRAIU, Benjamin REISER	
Nonparametric covariate adjustment for receiver operating characteristic curves.....	27
Luke BORNIN, Arnaud DOUCET, Raphael GOTTARDO	
An efficient computational approach for prior sensitivity analysis and cross-validation .....	47
Edit GOMBAY	
Change detection in linear regression with time series errors.....	65
Jean-Renaud PYCKE	
Some tests for uniformity of circular distributions powerful against multimodal alternatives .....	80
Ori DAVIDOV, Amir HERMAN	
Testing for order among $K$ populations: theory and examples .....	97
Azadeh MOGHTEADERI, Glen TAKAHARA, David J. THOMSON	
Unaliasing of aliased line component frequencies.....	116
Qingzhao YU, Bin LI, Zhide FANG, Lu PENG	
An adaptive sampling scheme guided by BART - with an application to predict processor performance .....	136
Paul D. MCNICHOLAS, T. Brendan MURPHY	
Model-based clustering of longitudinal data .....	153
Acknowledgement of referees' services Remerciements aux membres des jurys .....	169





# GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A pdf or paper copy may be required for formulas and figures.

## 1. Layout

- 1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The documents should be divided into numbered sections with suitable verbal titles.
- 1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

## 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

## 3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis.

## 4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

## 5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

## 6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

# DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de finaliser votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1.	<b>Présentation</b>	1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour. 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés. 1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte. 1.4 Les remerciements doivent paraître à la fin du texte. 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2.	<b>Résumé</b>	Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
3.	<b>Rédaction</b>	3.1 Éviter les notes au bas des pages, les abréviations et les sigles. 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc. 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin. 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique. 3.5 Distinguer clairement les caractères ambigus (comme w, $\omega$ ; o, O, 0 ; l, I). 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
4.	<b>Figures et tableaux</b>	4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
5.	<b>Bibliographie</b>	5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164). 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
6.	<b>Communications brèves</b>	6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.





## CONTENTS

## TABLE DES MATIÈRES

## Volume 38, No. 1, March/mars 2010

Paul GUSTAFSON	Report from the previous editor.....	1
Jiahua CHEN	Notes from the New Editor.....	5
Hung HUNG, Chin-Tsang CHIANG	Estimation methods for time-dependent AUC models with survival data.....	5
Fang YAO, Radu V. CRAIU, Benjamin REISER	Nonparametric covariate adjustment for receiver operating characteristic curves.....	27
Luke BORN, Arnaud DOUCET, Raphael GOTTARDO	An efficient computational approach for prior sensitivity analysis and cross-validation.....	47
Edi GOMBAY	Change detection in linear regression with time series errors.....	65
Jean-Renaud PYCKE	Some tests for uniformity of circular distributions powerful against multimodal alternatives.....	80
Or DAVIDOV, Amir HERMAN	Testing for order among $K$ populations: theory and examples.....	97
Azadeh MOGHTADERI, Glen TAKAHARA, David J. THOMSON	Unaliasing of aliased line component frequencies.....	116
Qingzhao YU, Bin LI, Zhide FANG, Lu PENG	An adaptive sampling scheme guided by BART - with an application to predict processor performance.....	136
Paul D. MCNICHOLAS, T. Brendan MURPHY	Model-based clustering of longitudinal data.....	153
	Acknowledgement of referees' services Remerciements aux membres des jurys.....	169

Julie ATHERTON, Benoit CHARBONNEAU, David B. WOLFSON, Lawrence JOSEPH, Xiaojie ZHOU, Alain C. VANDAL	Bayesian optimal design for changepoint problems	495
Jingjing WU, Rohana J. KARUNAMUNI	On minimum Hellinger distance estimation	514
Noomen Ben GHORBAL, Christian GENEST, Johanna NEŠLEHOVÁ	On the Ghoudi, Khoudraji, and Rivest test for extreme-value dependence	534
Christopher R. BILDER, Thomas M. LOUGHIN	Modeling multiple-response categorical data from complex surveys	553
Lajos HORVÁTH, Piotr KOKOSZKA, Matthew REIMHERR	Two sample inference in functional linear models	571
Jianguo SUN, Junshan SHEN	Efficient estimation for the proportional hazards model with competing risks and current status data	592
Haonan WANG, Jun ZHU	Variable selection in spatial regression via penalized least squares	607
Rui WANG, Stephen W. LAGAKOS	Inference after variable selection using restricted permutation methods	625
Liqun XI, Ray WATSON, Ji-Ping WANG, Paul S.F. YIP	Estimation in capture-recapture models when covariates are subject to measurement errors and missing data	645
Guosheng YIN, Hui LI	Least squares estimation of varying-coefficient hazard regression with application to breast cancer dose-intensity data	659

# JOURNAL OF OFFICIAL STATISTICS

## An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

### Contents

### Volume 25, No. 4, 2009

Innovative Statistics to Improve Our Notion of Reality	Henk K. van Tuinen .....	431
Research and Development in Official Statistics and Scientific Co-operation with Universities: A Follow-Up Study	Risto Lehtonen, Carl-Erik Särndal .....	467
Does the Effect of Incentive Payments on Survey Response Rates Differ by Income Support History?	Juan D. Barton, Robert V. Breunig, Deborah Cobb-Clark, Tue Gørgens, Anastasia Sarbayeva .....	483
Design of Web Questionnaires: The Effect of Layout in Rating Scales	Vera Toebe, Marcel Das, Arthur van Soest .....	509
The Effect of Single-Axis Sorting on the Estimation of a Linear Regression	Mathias Schmid .....	529
Using Bayesian Networks to Create Synthetic Data	Jim Young, Patrick Graham, Richard Penny .....	549
Evaluating Alternative One-Sided Coverage Intervals for a Proportion	Van K. Liu, Phillip S. Kott .....	569
Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB Establishment Survey	Jörg Drechsler, J.P. Reiter .....	589
Editorial Collaborators	.....	611

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)



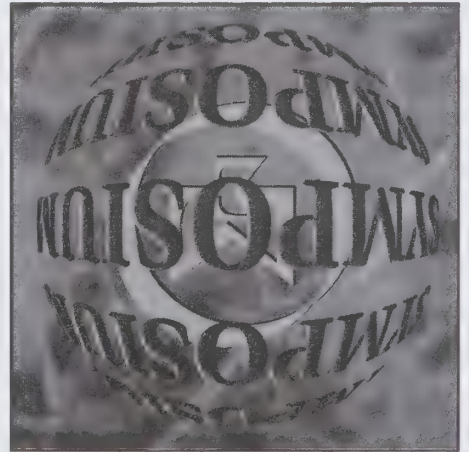
# ICES IV

## QUATRIÈME CONFÉRENCE INTERNATIONALE SUR LES ENQUÊTES-ÉTABLISSSEMENTS (ICES IV), 2012

Les préparatifs sont en cours en vue de la quatrième Conférence internationale sur les enquêtes-établissements (*International Conference on Establishment Surveys* (ICES)). Si vous avez déjà assisté à l'une de ces conférences, vous savez à quel point elles ont été utiles aux enquêtes-établissements, tant du point de vue de la documentation que de la pratique. Si vous êtes plutôt nouveau dans le domaine des enquêtes-établissements, vous trouverez cette conférence tout particulièrement enrichissante. Depuis la dernière ICES, tenue en 2007, un grand nombre de nouvelles techniques ont été mises au point par des praticiens du monde entier. Un des principaux attraits de la ICES est la forte présence de délégués internationaux, parmi les représentants et dans l'auditoire. Plus de 400 personnes provenant de 94 pays ont assisté à la ICES III. Du 11 au 14 juin 2012, des praticiens d'enquêtes provenant d'organismes gouvernementaux, du milieu universitaire, du secteur privé et d'ailleurs se réuniront au Centre Sheraton Montréal, dans la province de Québec, au Canada, à l'occasion de la ICES IV. Ils y poursuivront la tradition de mettre en commun leurs techniques novatrices et leurs meilleures pratiques pour traiter des questions présentant un intérêt pour tous.

Les réunions sont parrainées par l'*American Statistical Association* (ASA), la section de l'ASA sur les méthodes de recherche par enquête, la section de l'ASA sur les statistiques gouvernementales, l'Association internationale des statisticiens d'enquêtes et la Société statistique du Canada. Comme dans le passé, le soutien administratif de la ICES IV est assuré par l'ASA. De plus, de nombreux autres organismes gouvernementaux et organisations apporteront leur soutien à la conférence ou le font déjà.

Avec l'appui de tant d'organisations exemplaires et la présence de praticiens d'enquêtes-établissements d'horizons si divers, nous nous attendons à un autre rassemblement fructueux dans le cadre de cette série des ICES. Portez les dates du **11 au 14 juin 2012** à votre calendrier et venez rencontrer des praticiens du monde entier à **Montréal, au Canada** ! Vous pouvez participer au programme de la ICES IV, qui prend de l'ampleur, et discuter de questions d'actualité, de perspectives d'avenir et de méthodes de pointe touchant les enquêtes auprès des entreprises, des fermes et des établissements. De plus amples renseignements sur la participation et le programme vous seront communiqués au moyen de ce bulletin d'information et seront également affichés sur le site Web de la ICES IV sous peu. Les demandes de renseignements peuvent être envoyées à l'adresse [ices4@amstat.org](mailto:ices4@amstat.org).



## **Statistiques sociales : interaction entre recensements, enquêtes et données administratives**

Le Symposium international de 2010 sur les questions de méthodologie de Statistique Canada se déroulera du 26 au 29 octobre 2010 à l'hôtel Crowne Plaza (situé en plein cœur du centre-ville d'Ottawa).

Ce Symposium sera intitulé « **Statistiques sociales : interaction entre recensements, enquêtes et données administratives** ». Y sont conviés tous les membres de la communauté statistique, qu'ils proviennent d'organismes de recherche privés, gouvernementaux ou universitaires, et particulièrement ceux qui s'intéressent aux enjeux méthodologiques ou statistiques qui découlent de l'utilisation de sources multiples de données (recensements, enquêtes par sondage ou données administratives).

Des ateliers auront lieu le premier jour, tandis que des séances plénières et plusieurs séances de présentations en parallèle portant sur une vaste gamme de sujets se tiendront les jours suivants. D'autres recherches et résultats pourraient être communiqués au moyen de séances de présentations par affiches.

Les présentations porteront sur les aspects méthodologiques liés à l'utilisation de sources multiples de données. Les sujets de ces présentations pourraient inclure :

- Base et plan de sondage
- Coordination d'échantillons
- Contenu et conception de questionnaire
- Mode de collecte de données et obtention de données administratives
- L'utilisation de données administratives comme complément de données d'enquête
- Les données administratives pour l'estimation directe
- Les bases de données statistiques à partir de données administratives (ex. : les registres de population)
- Imputation
- Pondération et estimation
- Diffusion et accès aux données
- Techniques de couplage d'enregistrements
- Logiciels de couplage d'enregistrements
- Erreurs de mesure
- Fardeau de réponse
- Traitement de la non-réponse
- Enjeux liés à la confidentialité, à la vie privée et à l'éthique
- L'estimation sur petits domaines

Consultez notre site Web régulièrement, afin d'obtenir plus de détails sur le programme, les ateliers, l'inscription, l'hébergement, les activités touristiques et bien d'autres choses encore, à

<http://www.statcan.gc.ca/conferences/symposium2010/index-fra.htm>

## Bibliographie

- Mohadjer, L., et Curtin, L.R. (2008). Trouver l'équilibre entre les divers objectifs du plan d'échantillonnage de la National Health and Nutrition Examination Survey. *Techniques d'enquête*, 34, 1, 131-140.
- NCES (2001). Technical Report and Data File User's Manual For the 1992 National Adult Literacy Survey. U.S. Department of Education, National Center for Education Statistics, Washington, DC.
- OCDE (2005). Apprentissage et réussite - Premiers résultats de l'enquête sur la littératie et les compétences des adultes. Organisation de Coopération et de Développement Économiques, Paris. Statistique Canada, Ottawa.
- Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.
- Kalton, G., Brick, J.M. et Lè, T. (2005). Estimating Components of Design Effects for Use in Sample Design, Household Sample Surveys in Developing and Transition Countries, Chapitre VI United Nations, New York, 95-121.
- Cochran, W.G. (1977). *Sampling Techniques*. 3<sup>ème</sup> Ed. New York : John Wiley & Sons, Inc.
- Clark, R.G., et Steel, D.G. (2007). Sampling within Households in Household Surveys. *Journal of the Royal Statistical Society, Séries A*, 170, 63-82.



L'abandon de ménages donne essentiellement les mêmes résultats.

Remerciements

Les auteurs tiennent à souligner les contributions très utiles de Leyla Mohadjer et de Bob Fay.

Annexe A

Estimations de  $Rho^*$  selon la taille du ménage

Des estimations d'après les données d'enquête ne peuvent pas être obtenues pour  $Rho^*$  selon la taille du ménage, puisque 1 PE seulement a été sélectionnée quand la taille était égale ou inférieure à 3 et que la taille d'échantillon était trop faible pour produire des estimations pour chaque taille de ménage égale ou supérieure à 4. Par conséquent, les estimations de  $Rho^*$  selon la taille du ménage sont modélisées en utilisant des données de recensement. À la figure 2,  $Rho^*$  est représenté sur l'axe des y et la taille du ménage, sur l'axe des x. La courbe supérieure est celle obtenue pour l'échantillon du fichier de microdommées à grande diffusion (FMGD) du Recensement des États-Unis pour le niveau de scolarité pour les personnes de 25 ans et plus. Cette courbe supérieure montre que le niveau de scolarité est plus semblable pour les ménages comptant deux adultes, qui sont peut-être plus susceptibles d'être des couples mariés. Elle révèle une diminution lorsque l'on passe de 2 à 3 adultes dans le ménage. Nous avons saisi la variation de la taille des ménages en calculant le ratio de  $Rho^*$  pour les scores de compréhension de textes suivis de la NAAL à  $Rho$  pour le niveau de scolarité pour l'échantillon du FMGD du recensement parmi les ménages pour lesquels  $B > 3$  et en appliquant le ratio au  $Rho$  de

Annexe B

Algorithme de calcul

Nous avons élaboré un algorithme de calcul pour arriver aux tailles d'échantillon optimales dans les ménages pour chaque taille de ménage  $B$ . Nous avons construit l'algorithme de façon à générer des solutions optimales entières ou fractionnaires qui reflètent les effets de la mise en grappes, des taux d'échantillonnage différentiels et du coût, sous les contraintes qu'on sélectionne au moins une personne par ménage et pas plus de deux. Suivent les étapes de l'algorithme (toutes les exécutions du traitement convergent en quatre itérations) :

- Initialiser en fixant  $b = 1$  pour toutes les valeurs de  $B$  (Tirer1).
- Calculer  $DEFF_{MN}^{grappe}, DEFF_{pond}, c_p, c_{MN}$  et  $VC(0)$ .
- Faire  $I = 1$  à 5.
- Faire  $B = 1$  à 11.
- Calculer  $DEFF_{MN}^{grappe}, DEFF_{pond}, c_p, c_{MN}$  et  $VC(I)$  et  $VC(I-1)$ .
- Répéter les  $b_p$  ayant la valeur la plus faible de  $VC$ .
- Si  $VC(I) = VC(I-1)$ , s'arrêter.
- Fin.

Figure 2 Estimations de  $Rho^*$  pour la NAAL selon la taille du ménage

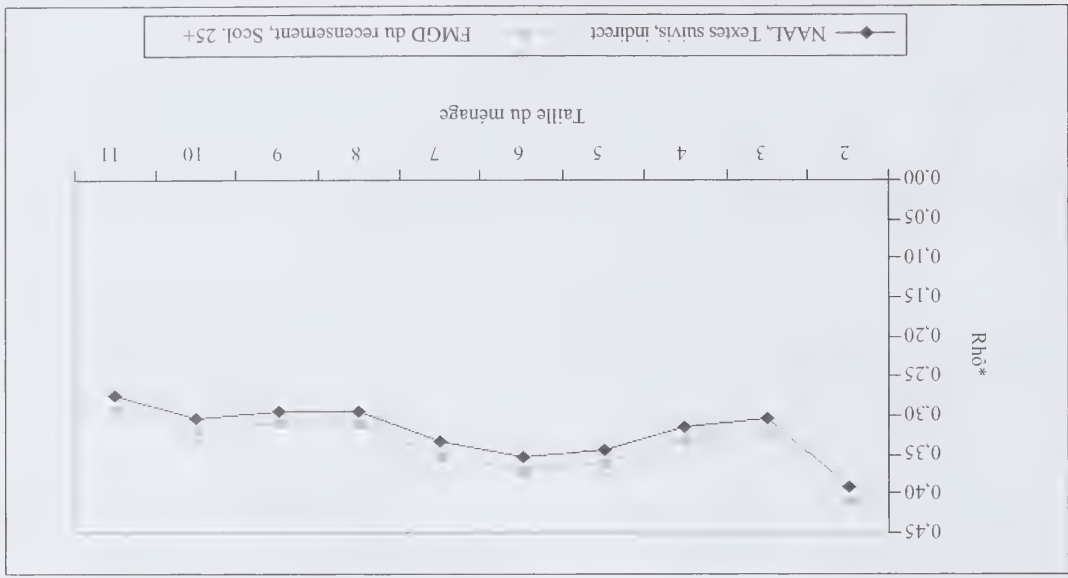


Tableau 6  
Nombre prévu optimal de personnes par ménage selon le type de méthode d'échantillonnage des personnes et la taille du ménage (B)

Méthode d'échantillonnage des personnes												
Nombre entier				Nombre fractionnaire				Abandon				
$k^*$	$C_{MN}/C_p$	$B = 1$	$B = 2$	$B = 3$	$B = 4$	$B = 1$	$B = 2$	$B = 3$	$B = 4$	$B = 1$	$B = 2$	$B = 3$
1	0,5	1	1	2	2	1,4	2	2	2	0,6	1,3	2
1	1	1	2	2	2	1,6	2	2	2	0,9	1,6	2
1	2	1	2	2	2	1,9	2	2	2	1	1,9	2
3	0,5	1	2	2	2	1,6	2	2	2	0,8	1,5	2
3	1	1	2	2	2	1,8	2	2	2	1	1,8	2
3	2	1	2	2	2	1	2	2	2	2	2	2

Tableau 7  
Réduction en pourcentage pour NAAL3 et les solutions optimales par rapport à la stratégie T1 et taille moyenne des grappes

Réduction en pourcentage par rapport à la stratégie T1												
Entier				Fractionnaire				Abandon				
$k^*$	$C_{MN}/C_p$	NAAL3	Entier	Fractionnaire	Abandon	NAAL3	Entier	Fractionnaire	Abandon	Fractionnaire	Abandon	Abandon
1	0,5	8,2	13,0	15,8	18,0	10,6	1,18	1,38	1,21	1,38	1,45	1,21
1	1	9,1	19,2	20,4	20,6	10,6	1,68	1,48	1,45	1,63	1,45	1,45
1	2	9,9	26,1	26,1	26,1	10,6	1,68	1,63	1,63	1,63	1,63	1,63
3	0,5	8,6	17,3	18,7	19,0	10,6	1,68	1,48	1,37	1,58	1,37	1,37
3	1	9,5	23,7	23,9	23,9	10,6	1,68	1,58	1,58	1,58	1,58	1,58
3	2	10,4	30,2	30,2	30,2	10,6	1,68	1,68	1,68	1,68	1,68	1,68

l'effet de plan dû à la mise en grappes à l'intérieur des ménages, tout en faisant varier la contribution des degrés de sélection des UPE et des USE à la variance totale dans les plans d'échantillonnage à plusieurs degrés. La mesure  $Rho^{**}$  est introduite en tant qu'expression de la corrélation intra-ménage sous un plan à plusieurs degrés en tenant compte de la contribution des deux premiers degrés d'échantillonnage à la variance totale.

En outre, nous avons élaboré un algorithme pour calculer les solutions optimales concernant la taille d'échantillon en intégrant les effets de plan dus à la mise en grappes, aux taux d'échantillonnage différentiels et aux coûts.

En général, les principaux déterminants de la réduction en pourcentage de la fonction VC par rapport à l'approche T1 sont le niveau de dominance des composantes de la variance dues aux UPE et aux USE dans l'échantillonnage à plusieurs degrés, le rapport des coûts et la règle de sélection utilisée. Pour la gamme de données évaluée,  $Rho^*$  a un effet limité sur la réduction de VC. Par rapport à l'approche T1, en général, la règle de la NAAL produit une amélioration par rapport à la stratégie très répandue de tirage d'une personne (T1er1). La règle d'un nombre entier optimal donne de meilleurs résultats que la règle de la NAAL. Cependant, la règle du nombre fractionnaire optimal n'offre qu'une amélioration limitée par rapport à la règle du nombre entier optimal. La règle de l'abandon optimal de ménages est meilleure que les autres règles pour les faibles rapports des coûts. Enfin, quand les deux premières composantes de la variance dominent et que le rapport des coûts est élevé, les règles du nombre entier, du nombre fractionnaire et de

### 5. Résumé

Enfin, nous avons effectué une analyse de sensibilité en faisant varier les valeurs de  $Rho^*$ . Nous avons ajusté un modèle de régression en fonction de la réduction en pourcentage, par rapport à la stratégie T1, de la fonction VC dans lequel les variables indépendantes étaient la stratégie utilisée (NAAL3, nombre entier, nombre fractionnaire, abandon de ménages), le rapport des coûts (0,1, 0,5, 1, 2, 10), le rapport des variances (1, 3, 5) et  $Rho^*$  (+/-0,1). Pour la gamme de données,  $Rho^*$  avait un effet limité (paramètre estimé de -7,4 avec une erreur-type associée de 4,5) sur la réduction en pourcentage de la fonction VC, tandis que les autres facteurs avaient un effet plus prononcé.

Plusieurs aspects du plan de sondage ont été pris en compte pour évaluer la règle de sélection dans les ménages pour la NAAL et l'ELCA, y compris les effets de grappes dus aux degrés initiaux d'échantillonnage. Afin de faciliter l'évaluation, nous avons formulé un moyen d'intégrer les contributions des UPE et des USE à la variance dans le calcul de l'effet de plan dû à la mise en grappes et à la correction intra-ménage pour décider du nombre de personnes et du nombre de ménages qu'il convient de sélectionner dans un plan d'échantillonnage à plusieurs degrés. Pour cela, nous introduisons la mesure de  $Rho^*$  compact, qui est calculée à l'intérieur de l'USE de sorte qu'il n'est pas influencé par les composantes de la variance dues aux UPE et aux USE. Cette approche est utile si l'on veut déterminer

Tableau 4  
Valeurs de  $k^*$  pour l'échantillon de l'ELCA

Estimation de l'ELCA
Score moyen – textes suivis
Score moyen – textes schématiques
Score moyen – textes à contenu quantitatif/numératif
Pourcentage au niveau 1 ou 2 – textes suivis
Pourcentage au niveau 1 ou 2 – textes schématiques
Pourcentage au niveau 1 ou 2 – numératif

Tableau 5  
Valeurs de  $Rho$ ,  $Rho^*$  et  $Rho^{**}$  pour les scores d'évaluation de la littérature

Estimation	NAAL	ELCA	NAAL	ELCA	$Rho^{**}$
Nombre de ménages avec 2 PE	930	162	930	162	0,95
Score moyen – textes suivis	0,57	0,33	0,60	0,38	0,19
Score moyen – textes schématiques	0,53	0,33	0,50	0,29	0,17
Score moyen – textes à contenu quantitatif/numératif	0,54	0,32	0,58	0,39	0,16
% sous le niveau de base (NAAL)/niveau 1 ou 2 (ELCA) – textes suivis	0,65	0,44	0,42	0,28	0,21
% sous le niveau de base (NAAL)/niveau 1 ou 2 (ELCA) – textes schématiques	0,61	0,37	0,39	0,28	0,20
% sous le niveau de base – textes quantitatifs (NAAL)/niveau 1 ou 2 (ELCA) numératif	0,62	0,36	0,40	0,17	0,20

4. Évaluation et résultats

Nous avons comparé les règles courantes d'échantillon-nage aux règles optimales d'échantillonnage en minimisant une fonction variance-coût (VC) qui est le produit des DEFF (c'est-à-dire, les accroissements de variance) dus à la formation de grappes et à la pondération ainsi que d'une fonction de coût qui est utilisée par Kish (1965) :

$$VC = DEFF_{MN}^{grappe} \times DEFF_{poid} \times n \left( c^p + \frac{b}{C_{MN}^p} \right)$$

où  $c^p$  = coût par personne ajoutée et  $C_{MN}^p$  représente le nombre de ménages échantillonnés. Afin de tenir compte des effets de grappe différents pour chaque taille de ménage  $B$ , nous remplaçons  $DEFF_{MN}^{grappe}$  par :

$$DEFF_{MN}^{grappe} = \frac{k^* + \sum_B \frac{M_B}{M} (1 + (b_B - 1) Rh\phi_B^*)}{k^* + 1}$$

où  $Rh\phi_B^*$  est calculé comme il est décrit à l'annexe A.

Notons que la fonction VC représente le coût additionnel de l'accroissement de la taille globale d'échantillon pour compenser l'accroissement de variance dû aux composantes du DEFF. Le tableau 6 donne les résultats pour les solutions entières optimales calculées au moyen d'un algorithme qui est décrit à l'annexe B. Le tableau montre que, quand le ratio des coûts passe de 0,5 à 1 pour  $k^* = 1$ , nous souhaiterions tirer plus de personnes par ménage, c'est-à-dire 2 sur 2 au lieu de 1 sur 2. Quand le ratio des variances passe de 1 à 3 pour les solutions entières optimales, un changement n'est observé que pour la taille de ménage de 2 et le rapport des coûts de 0,5. Autrement dit, quand le rapport des variances est égal à 3, il est avantageux de tirer 2 personnes sur 2 au lieu de 1 personne sur 2.

Le tableau 6 donne les résultats quand des tailles fractionnaires d'échantillon sont permises. Les rapports des variances et des coûts pour la NAAL et l'ELCA ont tendance à être de l'ordre de 1, où il semble que la sélection de 1 personne sur 1, de 1,6 personne sur 2 et de 2 personnes autrement est la meilleure règle. Les effets des rapports des coûts et des variances sont plus clairs sous le scénario des tailles d'échantillon fractionnaires que sous celui des solutions entières.

Si le coût de l'exécution d'une présélection est faible comparativement à celui de l'interview, les variances peuvent être réduites en utilisant l'approche fractionnaire d'abandon de ménages. Le tableau 6 donne les tailles d'échantillon avec abandon optimales. Selon cette approche, par exemple, une taille d'échantillon de 0,9 indique que nous abandonnons 10 % des ménages, où  $B = 1$ . Si le coût de la présélection est une part très faible du coût d'interview, le plan d'échantillonnage optimal pourrait comporter l'abandon d'un beaucoup plus grand nombre de ménages.

Sous les paramètres NAAL/ELCA probables pour les ratios des coûts ( $C_{MN}^p / C^p = 1$ ) et les ratios des variances ( $k^* = 1$ ), quand elle est comparée à l'approche Tivert, la fonction VC peut être réduite d'environ 9 % en utilisant la règle d'échantillonnage de la NAAL/ELCA, 19 % en utilisant la solution entière optimale et 20,6 % en utilisant l'approche avec abandon de ménage optimale. En général, les gains dus à l'écart par rapport à l'approche Tivert augmentent à mesure que le coût par ménage supplémentaire (c'est-à-dire de la présélection) augmente. Les tailles moyennes de grappes pour chaque approche sont données au tableau 7. Pour la NAAL et la règle entière optimale, la taille moyenne de grappe indique le pourcentage de ménages avec 2 PE. Par exemple, environ 6 % des ménages compteraient 2 PE sous la stratégie NAAL3.





*Composition du ménage.* Enfin, on pourrait prendre en considération la composition du ménage et les relations entre les personnes dans un ménage pour concevoir la règle de sélection. Le tableau 3 donne les valeurs de Rhô pour diverses relations entre les membres du ménage, pour un ménage avec 2 PE dans l'enquête NAAL. Rhô varie fortement selon les relations entre les membres du ménage. Ces relations ont été déterminées d'après le sexe et l'âge.

### 3. Estimation de Rhô intra-ménage et de DEFF sous échantillonnage à plusieurs degrés

Jusqu'à présent, la discussion au sujet de Rhô se rapportait à un plan d'échantillonnage à deux degrés, mais la NAAL ainsi que l'ELCA ont toutes deux un plan d'échantillonnage à quatre degrés. La variance totale peut être décomposée en quatre termes de variance de type inter attribuables aux UPE, aux USE, aux ménages et aux personnes, de la façon suivante :

$$\sigma^2_T = \sigma^2_{UPE} + \sigma^2_{USE(UPE)} + \sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)}.$$

Comme nous le montrons plus bas, si nous suivons l'approche utilisée pour l'échantillonnage à deux degrés pour estimer Rhô pour un plan d'échantillonnage à quatre degrés, le numérateur contient non seulement la composante de variance inter-ménages, mais également les contributions des composantes inter-UPE et inter-USE qui accroissent les valeurs de Rhô.

où  $n_{UPE}$ ,  $n_{USE}$ ,  $n_{MN}$  et  $bn_{MN}$  représentent respectivement les tailles d'échantillon des UPE, des USE, des ménages et des personnes.

$$Var(\hat{\theta}) = \frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}} + \frac{\sigma^2_{MN(USE)}}{n_{MN}} + \frac{\sigma^2_{PERS(MN)}}{bn_{MN}}.$$

En utilisant le Rhô\* compact, nous calculons maintenant le DEFF estimé sous un plan d'échantillonnage à plusieurs degrés afin de déterminer la taille optimale d'échantillon dans les ménages. La variance d'une estimation ( $\hat{\theta}$ ) avec  $b$  personnes par ménage peut être décomposée comme il suit :

$$Rh\hat{o}^* = \frac{\sigma^2_{MN(USE)}}{\sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)}}.$$

Par conséquent, pour évaluer les règles d'échantillonnage dans les ménages sous un plan à plusieurs degrés, nous supposons que le plan d'échantillonnage des UPE et des USE sera le même dans l'avenir. Nous pouvons accomplir cela en nous limitant à considérer l'échantillonnage à l'intérieur des USE. Par conséquent, le calcul de Rhô est contenu dans les USE, c'est-à-dire qu'il est effectué de manière compacte, sans effet dû aux composantes UPE et USE. Nous parlons dans ce cas du Rhô compact (c'est-à-dire dans les USE) dénoté par  $Rh\hat{o}^*$ , exprimé par :

$$Rh\hat{o} = 1 - \frac{\sigma^2_{PERS(MN)}}{\sigma^2_{UPE} + \sigma^2_{USE(UPE)} + \sigma^2_{MN(USE)}}.$$

Tableau 3  
Rhô pour les scores d'évaluation de la NAAL selon la relation entre les membres du ménage

Estimation	Frères et sœurs	Enfant-tuteur	Marées	Autre
Nombre de ménages avec 2 PE	111	205	180	434
Score moyen – textes suivis	0,42	0,35	0,70	0,59
Score moyen – textes schématiques	0,40	0,27	0,72	0,54
Score moyen – textes à contenu quantitatif	0,46	0,36	0,63	0,56
Pourcentage sous le niveau de base – textes suivis	0,52	0,41	0,79	0,67
Pourcentage sous le niveau de base – textes schématiques	0,54	0,40	0,78	0,60
Pourcentage sous le niveau de base – textes à contenu quantitatif	0,51	0,41	0,77	0,65

*Considérations budgétaires.* Le coût de la présélection d'un ménage dans le cas d'un plan d'échantillonnage de 1 PE par ménage comparativement au coût de l'interview/évaluation d'une deuxième personne dans un ménage est examiné dans une analyse détaillée présentée plus loin.

*Systèmes informatisés.* Les systèmes informatisés, tels que l'interview sur place assistée par ordinateur (IPAO), ont la capacité de traiter des tailles d'échantillon fractionnaires. En d'autres mots, il est possible de programmer la sélection aléatoire de 1 ou 2 PE étant donné une taille d'échantillon fractionnaire préattribuée. Les systèmes informatisés ont également la capacité de trier la liste de personnes admissibles et de sélectionner 2 PE selon un échantillonnage aléatoire systématique. Un autre avantage tient au fait que le programme de sélection peut être testé et validé avant la collecte des données.

*Domaines d'intérêt.* Comme nous l'avons mentionné plus haut, l'échantillonnage optimal dans les ménages dépend de l'importance de l'effet de grappe associé à la variable d'intérêt. L'effet de grappe peut être plus faible quand la variable est associée à un sous-groupe de la population plutôt qu'à la population complète. Par exemple, quand un domaine de déclaration clé est le sexe dans une enquête auprès de la population adulte, la catégorie des hommes comportera vraisemblablement la sélection de 1 PE par ménage en moyenne et sera moins susceptible de comporter la sélection de 2 PE de sexe masculin qui introduit un effet de grappe. Par conséquent, lorsqu'il existe de multiples domaines d'intérêt dans un ménage typique, il est souvent avantageux de sélectionner plus de 1 PE dans un ménage. Voir Mohadjer et Curtin (2008) pour un exemple de considérations concernant le plan d'échantillonnage pour des enquêtes axées sur de multiples sous-groupes de la population.

aucune considération des coûts. Nous avons utilisé pour cet exemple la distribution nationale des ménages américains selon la taille établie d'après la Current Population Survey de 2007. Comme l'illustre la figure 1, l'approche fractionnaire est la meilleure règle pour une grande gamme de valeurs de  $Rho$ . L'approche fractionnaire peut être programmée dans un système informatisé lorsque l'on dénombre et sélectionne les membres des ménages (une discussion plus approfondie des systèmes informatisés suit). Si aucun système informatisé n'est disponible pour la présélection, la meilleure approche pour les faibles valeurs de  $Rho$  est celle où la mise en grappes est plus importante, c'est-à-dire la règle 2, tandis que la règle NAAL3 est la meilleure pour les valeurs de  $Rho$  plus grandes qu'environ 0,34.

*Echantillonnage à plusieurs degrés.* Pour les plans d'échantillonnage aréolaire à plusieurs degrés, l'effet de la mise en grappes sur l'échantillonnage dans les ménages est affecté par la mise en grappes due aux UPE et aux USE. Comme l'a fait remarquer Kish (1965), la mise en grappes des ménages et des personnes dans les UPE et dans les USE accroît la variance d'échantillonnage (autrement dit, les unités dans les UPE et dans les USE sont plus semblables les unes aux autres). L'effet marginal de la mise en grappes dans les ménages peut être atténué par la domination des composantes de la variance liées aux UPE et aux USE (cependant, la grandeur de l'effet peut différer selon le type d'estimations et de variables). Autrement dit, un plus grand nombre de personnes peuvent être sélectionnées dans un ménage pour des enquêtes où l'importance de la mise en grappes est grande à cause des deux premiers degrés d'échantillonnage. Des précisions concernant cette distinction sont fournies à la section 3.

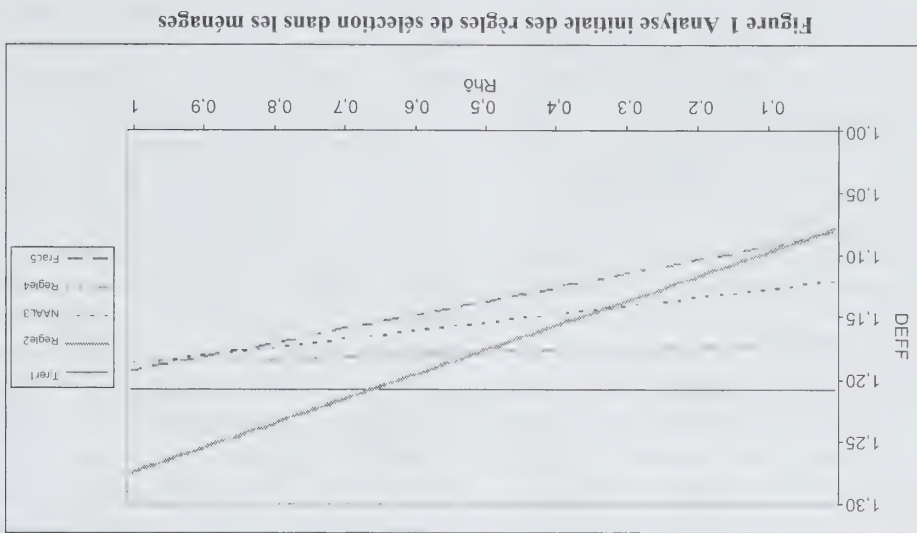


Figure 1 Analyse initiale des règles de sélection dans les ménages



*Mise en grappes des personnes dans les ménages.* Kish (1965) discute des avantages d'un échantillon en grappes comparativement à un échantillon aléatoire simple. Habituellement, le coût par personne est plus faible pour un échantillon en grappes, mais la variance unitaire est plus élevée et pose de plus grandes difficultés dans l'analyse statistique. Kish a introduit le concept d'un effet de plan (DEFF), qui mesure l'accroissement de la variance dû aux écarts par rapport à un échantillon aléatoire simple, tels que la mise en grappes des personnes dans les ménages. Dans de nombreuses enquêtes, la sélection est limitée à une personne échantillonnée (PE) par ménage, en raison de préoccu-pations quant à l'accroissement de l'effet de grappe (c'est-à-dire l'accroissement de l'effet sur les estimations de la variance) associé à la sélection de plusieurs PE par ménage. L'effet de plan dû à la mise en grappes peut être exprimé sous la forme :  $DEFF_{grappe} = 1 + (b - 1) Rhô$ , où  $b = \sum (M_b / M) b_b$ ,  $M_b$  = nombre de ménages de taille  $b$ ,  $M$  = nombre de ménages, et  $b_b$  = taille de l'échantillon de personnes dans les ménages de taille  $b$  (Kish 1965). Cette composante de l'effet de plan augmente quand la taille de l'échantillon dans un ménage augmente ou que la valeur de la corrélation intra-grappe ( $Rhô$ ) augmente. Comme dans Cochran (1977),  $Rhô$  peut être approximé par

$$Rhô = 1 - \frac{\sigma_w^2}{\sigma^2}$$

ou

$$\sigma_w^2 = \sum_a \sum_b^{j=1} (y_{aj} - \bar{y}_{.j})^2 / (n - a)$$

et

$$\sigma^2 = \sum_a \sum_b^{j=1} (y_{aj} - \bar{y}_{..})^2 / (n - 1),$$

*Taux d'échantillonnage différentiels.* L'effet de grappe n'est pas le seul facteur qui accroît la variance. Celle-ci peut également augmenter à cause des taux d'échantillonnage différentiels (qui donnent lieu à une pondération différentielle). Sous la stratégie de 1 PE par ménage, l'accroissement est directement associé à la variation de la taille du ménage, puisque le taux d'échantillonnage pourrait varier de 1 sur 1 à 1 sur 7 ou plus. L'effet de plan dû aux taux d'échantillonnage différentiels est exprimé sous la forme  $DEFF_{pond} = \sum (p_b / k_b) \sum (p_b k_b)$ , où  $p_b = N_b / N$ ,  $N_b$  = nombre de personnes admissibles dans la population dans les ménages de taille  $b$ ,  $N$  = nombre de personnes

1. Tirer1 : 1 PE quelle que soit la taille du ménage.
2. Règle2 : 1 PE pour les tailles de ménages jusqu'à 2 ; sinon 2 PE sélectionnées.
3. NAAL3 : 1 PE pour les tailles de ménages allant jusqu'à 3 ; sinon 2 PE sélectionnées.
4. Règle4 : 1 PE pour les tailles de ménages allant jusqu'à 4 ; sinon 2 PE sélectionnées.
5. Frac5 : Tirer au moins 1 PE, mais pas plus de 2 PE et si la taille d'échantillon pour un ménage contenant deux personnes admissibles est 1,6, deux personnes sont sélectionnées 60 % du temps au hasard, et une personne est sélectionnée 40 % du temps.

Bien que l'approche Tirer1 n'essaye pas de réduire l'effet de plan dû aux taux d'échantillonnage différentiels, elle ne comporte pas d'effet de grappe. Cependant, les quatre autres approches susmentionnées donnent lieu à une réduction de la composante due aux taux d'échantillonnage différentiels, tout en introduisant un effet de grappe. Dans le cas de Frac5, sous l'hypothèse que les poids  $\pi$  sont utilisés, comme nous le supposons tout au long du présent exposé, l'approche sera celle donnant la réduction la plus importante de la composante due aux taux d'échantillonnage différentiels. L'ap-proche de poids proportionnels (ratio), le taux d'échantill-onnage différentiel augmente et l'avantage est moins clair et dépend de  $Rhô$ . La figure 1 illustre les meilleures options sous un plan d'échantillonnage des ménages à deux degrés avec taille fixe de l'échantillon effectif de personnes, sans

admissibles dans la population, et  $k_b$  = taux d'échantill-onnage dans les ménages de taille  $b$  (Kish 1965). Sous certaines conditions, l'effet de plan global peut être exprimé comme étant le produit des composantes de mise en grappes et de taux d'échantillonnage différentiels, soit :  $DEFF = DEFF_{grappe} \times DEFF_{pond}$ . Kalton, Brick et Lé (2005) laissent entendre que ce produit est applicable quand les poids sont aléatoires ou approximativement aléatoires.

Pour arriver à un échantillon autopondéré, les personnes pourraient être sélectionnées dans les ménages à un taux constant. Cependant, dans la plupart des enquêtes, la pré-férence n'est pas donnée à l'approche fondée sur les taux, parce qu'elle aboutirait à abandonner une partie des ménages ne comportant qu'une seule personne et, donc, augmenterait le coût de l'enquête. Nous limitons les diverses règles prises en considération à celles avec un minimum de 1 PE par ménage. Afin de ne pas imposer un trop lourd fardeau aux ménages, la taille maximale d'échan-tillon a été fixée à deux. Les règles d'échantillonnage envisagées sont les suivantes :

**2. Considérations concernant le plan d'échantillonnage**

Un certain nombre de facteurs doivent être pris en considération pour évaluer les règles de sélection dans les ménages pour des enquêtes telles que la NAAL et l'ELCA. La suite de la présente section porte sur l'effet qu'ont sur l'échantillonnage dans les ménages le fardeau de réponse, la mise en grappes des personnes dans les ménages, les taux d'échantillonnage différentiels, l'échantillonnage à plusieurs degrés, les considérations budgétaires, les systèmes informatisés, les domaines d'intérêt et la composition du ménage.

*Fardeau de réponse du ménage.* Dans les enquêtes sur l'alphabétisation des adultes, l'exécution de l'interview et de l'évaluation prend, en tout, environ une heure et demie. Par conséquent, l'une des préoccupations si l'on sélectionne plus d'une personne par ménage est l'accroissement du fardeau de réponse du ménage et l'effet sur les taux de réponse. Toutefois, comme le montre le tableau 2, l'écart entre les taux de refus de participer à l'ELCA et à la NAAL observés pour les ménages à 1 PE et à 2 PE n'est pas significatif (seuil de signification de 0,05).

Statistique Canada. En 2003, l'échantillon des États-Unis, commandité par le NCES, faisait partie de l'étude comparative destinée à mesurer les compétences des adultes dans plusieurs pays. Comme la NAAL, l'ELCA a été réalisée auprès d'un échantillon en grappes à plusieurs degrés et mesurait la compréhension de textes suivis et de textes schématisés, ainsi que la numération (OCDE 2005). L'échantillon de la NAAL était beaucoup plus grand (18 500 évaluations achevées) que celui de l'ELCA (3 400 évaluations achevées), et la population cible de la NAAL comprenait les personnes de 16 ans et plus, tandis que celle de l'ELCA comprenait les personnes de 16 à 65 ans. Le tableau 1 résume le plan et la structure de chaque enquête.

À la section 2, nous présentons une discussion des aspects du plan de sondage pris en considération qui ont aidé à formuler l'évaluation des règles d'échantillonnage dans les ménages. À la section 3, nous discutons du calcul des corrélations intra-ménage sous les plans d'échantillonnage à plusieurs degrés et nous nous concentrons sur l'intégration de l'effet de grappe résultant des premiers degrés de sélection de l'échantillon pour décider d'une règle de sélection dans les ménages. À la section 4, nous procédons à une évaluation des règles de sélection en utilisant des données provenant d'enquêtes sur la littératie des adultes menées sur place et présentons les résultats. Enfin, à la section 5, nous résumons brièvement l'étude.

**Tableau 1**  
Caractéristiques de la NAAL et de l'ELCA

Enquête	Echantillon aréolaire	Évaluations complètes	Collecte des données	Évaluations	Âge	Règle d'échantillonnage dans les MN
NAAL	UPE, USE, ménages, personnes	18 500	Présélection	Interview	16 ans et plus	$B \leq 3, b = 1$ $B > 3, b = 2$
ELCA	UPE, USE, ménages, personnes	3 400	Présélection	Interview	16 à 65 ans	$B \leq 3, b = 1$ $B > 3, b = 2$
Nota : UPE = Unité primaire d'échantillonnage, USE = Unité secondaire d'échantillonnage, $b$ = taille de l'échantillon, $B$ = taille du ménage.						

**Tableau 2**  
Taux de refus par les ménages à 1 PE et 2 PE pour les enquêtes sur la littératie des adultes

Enquête	Sous-groupe	Taux de refus
NAAL	Ménages à 1 PE	16,3
	Ménages à 2 PE	15,7
ELCA	Ménages à 1 PE	17,6
	Ménages à 2 PE	16,2

Nota : PE = personne échantillonnée.



# Evaluation des règles de sélection dans les ménages sous un plan à plusieurs degrés

Tom Krenzke, Lin Li et Keith Rust

## Résumé

La National Assessment of Adult Literacy (NAAL) de 2003 et l'Enquête internationale sur la littératie et les compétences des adultes (EILCA) comportaient chacune un plan d'échantillonnage aréolaire stratifié à plusieurs degrés. Le dernier degré consistait à dresser la liste des membres du ménage, à déterminer la situation d'admissibilité de chaque individu et à appeler la procédure de sélection pour sélectionner aléatoirement une ou deux personnes admissibles dans le ménage. L'objectif du présent article est d'évaluer les règles de sélection dans les ménages sous un plan d'échantillonnage à plusieurs degrés en vue d'améliorer la procédure dans de futures enquêtes sur la littératie. L'analyse est fondée sur la distribution courante des ménages américains selon leur taille et sur les coefficients de corrélation intra-grappe en utilisant les données sur la littératie en grappes, des taux d'échantillonnage différents, du coût par interview et du fardeau de réponse au niveau du ménage. Dans ce contexte, nous étendons une évaluation de l'échantillonnage dans les ménages sous un plan à deux degrés à un plan à quatre degrés et nous procédons à certaines généralisations aux échantillons à plusieurs degrés pour divers rapports de coûts.

Mots clés : Corrélation intra-grappe ; effets de plan ; échantillonnage à plusieurs degrés.

## 1. Introduction

La National Assessment of Adult Literacy (NAAL) de 2003, réalisée par le National Center for Education Statistics, a fourni aux chercheurs, aux praticiens, aux décideurs et au grand public un indicateur du progrès de la nation en ce qui concerne la maîtrise de l'anglais. Comme dans la National Adult Literacy Study (NALS) de 1992, on a évalué la compréhension de textes suivis, de textes schématiques et de textes au contenu quantitatif par les adultes présents dans les ménages. La conception des livrets de réponse était basée sur celle de la NALS de 1992 pour permettre la mesure des tendances entre 1992 et 2003.

Afin de réduire le coût des déplacements des intervieweurs pour rendre visite aux ménages, la NAAL comporte un plan d'échantillonnage en grappes stratifié à quatre degrés qui a produit 18 500 évaluations complètes administrées à des adultes de 16 ans et plus. Dans la NAAL, les comités ont été groupés pour former des unités primaires d'échantillonnage (UPB), qui ont été stratifiées et sélectionnées au premier degré. Au deuxième degré, des unités secondaires d'échantillonnage (USE) ont été formées et sélectionnées parmi les UPB échantillonnées. Les USE correspondaient à des îlots de recensement individuels ou à des groupes d'îlots adjacents comptant au moins 60 ménages (MN) qui ont été formés à l'intérieur des limites des secteurs. Subséquentement, les ménages ont été sélectionnés parmi les USE, et une personne échantillonnée (1 PE) a été sélectionnée aléatoirement dans les ménages dont la taille était égale ou inférieure à 3 ( $B \leq 3$ ) et deux personnes (2 PE) ont été sélectionnées dans les ménages dont la taille

était supérieure à 3 ( $B > 3$ ), où  $B$  désigne le nombre de personnes admissibles par ménage. Cette règle suit l'approche d'échantillonnage dans les ménages utilisée au premier cycle de la NAAL (NCES 2001), réalisée en 1992. Une évaluation de la règle de sélection a été effectuée en utilisant la distribution courante des ménages américains selon la taille et les coefficients de corrélation intra-classe calculés d'après les données de l'enquête de 2003. Pour cela, nous avons étendu une évaluation de l'échantillonnage dans les ménages sous un plan d'échantillonnage à deux degrés (Clark et Steel 2007) à un plan d'échantillonnage à quatre degrés, tel que celui utilisé dans l'enquête NAAL et avons fait certaines généralisations aux échantillons à plusieurs degrés pour divers ratios de coûts.

Les données utilisées pour l'évaluation comprennent les mesures de la littératie faites sur trois échelles correspondant à trois types de littératie, à savoir la compréhension de textes suivis, la compréhension de textes schématiques et la compréhension de textes au contenu quantitatif. Pour plus de renseignements au sujet des types de littératie de la NAAL, consulter [http://nces.ed.gov/NAAL/fr\\_tasks.asp](http://nces.ed.gov/NAAL/fr_tasks.asp). Deux types d'estimations sont utilisés, à savoir les moyennes (par exemple, score moyen de compréhension de textes suivis) et le pourcentage d'adultes à un certain niveau de littératie (par exemple, pourcentage sous le niveau de base pour la compréhension de textes suivis). Pour une discussion des niveaux de littératie utilisés dans la NAAL, consulter [http://nces.ed.gov/NAAL/perf\\_levels.asp](http://nces.ed.gov/NAAL/perf_levels.asp). En plus des données de la NAAL, l'évaluation s'appuie sur les données de l'échantillon des États-Unis de l'Enquête sur la littératie et les compétences des adultes (EILCA), qui a été menée par



- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press.
- Nathan, G. (2001). Méthodes de téléenquêtes applicables aux enquêtes-ménages - Revue et réflexions sur l'avenir. *Techniques d'enquête*, 27-1, 7-34.
- Rao, J.N.K., et Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- Sharp, J.S., et Adua, L. (2009). The social basis of agro-environmental concern: Physical versus social proximity. *Rural Sociology*.
- Sharp, J.S., et Clark, J.K. (2008). Between the country and the concrete: Rediscovering the rural-urban fringe. *City & Community*, 7-1, 61-79.
- Shettle, C., et Mooney, G. (1999). Monetary incentives in U.S. government surveys. *Journal of Official Statistics*, 15-2, 231-250.
- Singer, E., Van Hoewyk, J. et Maher, M.P. (1998). Does the payment of incentives create expectation effects? *The Public Opinion Quarterly*, 62-2, 152-164.
- Singer, E., Van Hoewyk, J. et Maher, M.P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64, 171-188.
- Singer, E. (2006). Introduction: Nonresponse bias in household surveys. *Public Opinion Quarterly*, 70-5, 637-645.
- Teitler, J.O., Reichman, N.E. et Sprachman, S. (2003). Costs and benefits of improving response rate for a hard-to-reach population. *Public Opinion Quarterly*, 67, 126-138.
- Trussell, N., et Lavrakas, P.J. (2004). The influence of incremental increases in token cash incentives on mail survey response. *Public Opinion Quarterly*, 68-3, 349-367.
- Visser, P.S., Krosnick, J.S., Marquette, J. et Curtin, M. (1996). Mail surveys for election forecasting? An evaluation of the columbus dispatch poll. *Public Opinion Quarterly*, 60, 181-227.
- Weisberg, H.F. (2005). *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago and London : The University of Chicago Press.
- Williams, R. (2006). Generalized ordered logit/Partial proportional odds models for ordinal dependent variables. *The Stata Journal*, 6-1, 58-82.
- Yammarino, F.J., Skinner, S.J. et Childers, T.L. (1991). Understanding mail survey response behavior: A meta-analysis. *Public Opinion Quarterly*, 55-4, 613-639.
- Yu, J., et Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research*, 20, 36-44.
- Lobao, L.M. (1990). *Locality and Inequality: Farm and Industry Structure and Socioeconomic Conditions*. Albany, NY : The State University of New York Press.
- Keeter, S., Miller, C., Kohut, A., Groves, R.M. et Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-48.
- Kaldenberg, D.O., Koenig, H.S. et Becker, B.W. (1994). Mail survey response patterns in a population of the elderly. *Public Opinion Quarterly*, 58-1, 68-76.
- James, J.M., et Bolstein, R. (1992). Large monetary incentives and their effect on mail survey response rates. *Public Opinion Quarterly*, 56, 442-453.
- Hansen, R.A. (1980). A self-perception interpretation of the effects of monetary and nonmonetary incentives on mail survey response behavior. *Journal of Marketing Research*, 17, 77-83.
- Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70-5, 646-675.
- Groves, R.M., Couper, M.P., Presser, S., Singer, E., Tourangeau, R., Acosta, G.P. et Nelson, L. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly*, 70-5, 139-147.
- Groves, R.M., Fowler, Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E. et Tourangeau, R. (2004). *Survey Methodology*. Hoboken, NJ : John Wiley & Sons, Inc.
- Groves, R.M., Singer, E. et Comins, A. (2000). Leverage-saliency theory of survey participation. *Public Opinion Quarterly*, 64, 299-308.
- Groves, R.M., et Cooper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York : John Wiley & Sons, Inc.
- Goyder, J.C. (1982). Further evidence of factors affecting response rates to mailed questionnaires. *American Sociological Review*, 47, 550-553.
- Gouldner, A. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25-2, 161-178.
- Fuchs, C., et Kenett, R. (2007). Missing data and imputation. Dans *Encyclopedia of Statistics in Quality and Reliability*. (Eds., F. Ruggeri, R.S. Kenett et F. Faltin). Wiley, 1090-1099.
- Freudenburg, W.R. (1991). Rural-urban differences in environmental concern: A closer look. *Sociological Inquiry*, 61-2, 167-198.
- Fox, R.J., Crask, M.R. et Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *The Public Opinion Quarterly*, 52-4, 467-491.
- Fowler, Jr. F.J. (2002). *Survey Research Methods*. Troisième Edition. Thousand Oaks, CA : Sage.
- Dunlap, R.E., et Hefterman, R.B. (1975). Outdoor recreation and environmental concern: An empirical examination. *Rural Sociology*, 40, 18-30.

variables de contrôle dans la présente étude, le fait d'avoir constaté qu'ils étaient liés significativement à la non-réponse partielle soulève certaines préoccupations pratiques quant au traitement des données manquantes dans les données d'enquête. Avant de choisir parmi diverses techniques de traitement des données manquantes (voir Fuchs et Kenett 2007), les analystes doivent vérifier s'il existe un biais éventuel de non-réponse résultant des effets de ces variables, surtout si elles font partie d'une analyse.

### Remerciements

Nous remercions Dr Hebert Weisberg du Political Science Department de l'Ohio State University (OSU) pour ses commentaires d'une version précédente de cet article. L'enquête sur laquelle celui-ci est basé a été subventionnée par l'Ohio Agricultural Research and Development Center (OARDC) au College of Food, Agriculture and Environment mental Sciences à l'OSU. Ces remerciements n'excluent pas le fait que nous sommes les seuls responsables du contenu de l'article.

### Bibliographie

Audirac, I. (1999). Unsettled views about the fringe: Rural-urban or urban-rural frontiers. Dans *Contested Countryside: The Rural Urban Fringe in North America*, (Eds., O.J. Furuseth et M.B. Lapping). Brookfield, VT : Ashgate, 7-32.

Brehm, J. (1993). *The Phantom Respondent*. Ann Arbor : University of Michigan Press.

Cameron, A.C., et Trivedi, P.K. (2009). *Microeconometrics using Stata*. College Station, Texas : Stata Press.

Church, A.H. (1993). Estimating the effects of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57, 62-79.

Curtin, R., Presser, S. et Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413-28.

Curtin, R., Presser, S. et Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Survey*, 69-1, 87-98.

Davern, M., Rockwood, T.H., Shertod, R. et Campbell, S. (2003). Repaid monetary incentives and item nonresponse in face-to-face interviews. *Public Opinion Quarterly*, 67, 139-147.

Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York : John Wiley & Sons, Inc.

Dillman, D.A. (2000). *Mail and Internet Surveys: The Tailored Design Method*. New York : John Wiley & Sons, Inc.

Dillman, D.A., Eltinge, J., Groves, R. et Little, R. (2002). Survey nonresponse in design, data collection, and analysis. Dans *Survey nonresponse*, (Eds., Groves et coll.). New York : John Wiley & Sons, Inc.

participer à l'enquête que ceux qui n'en avaient pas reçu, après l'introduction de contrôles pour d'autres variables dans le modèle de régression logistique. L'analyse démontre que la suppression complète des primes d'incitation réduit la probabilité de participer, indépendamment du contexte résidentiel du répondant. Même si nous n'avons pas relié ouvertement les primes d'incitation préparées à la théorie du levier et de la saillance de Groves et coll. (2000) dans les sections précédentes de notre discussion pour simplifier l'analyse, nos résultats apportent un soutien empirique supplémentaire à cette théorie. Ils montrent clairement que la prime d'incitation financière symbolique incluse dans chaque trousse d'enquête a contribué à l'accroissement de la participation dans les régions métropolitaines ainsi que non métropolitaines de l'Ohio, bien que l'effet ait été plus prononcé dans les premières. Ce résultat offre une nouvelle justification de l'usage répandu des primes d'incitation pour accroître les taux de réponse. Comme nous l'avons mentionné plus haut dans le présent article, cet usage répandu des primes d'incitation préparées dans les enquêtes rend nécessaire l'évaluation périodique de l'utilité de cette pratique. Les résultats laissent également entendre qu'il est nécessaire de vérifier la présence d'un biais éventuel de réponse si les primes d'incitation sont fournies uniquement à une partie des répondants échantillonnés, comme dans les cas où les primes d'incitation préparées ont pour cibles les personnes considérées comme étant moins susceptibles de participer.

En ce qui concerne la relation entre les primes d'incitation et la non-réponse partielle, nous ne trouvons aucune variation significative du nombre de points de données manquants entre les répondants qui ont reçu une prime d'incitation financière et ceux qui n'en ont pas reçu, contrairement à notre cinquième hypothèse. Ce résultat, obtenu en neutralisant les effets du lieu de résidence (proximité par rapport au paysage agricole et rural) et d'autres variables pertinentes, concorde avec ceux de travaux antérieurs de Davern et coll. (2003), qui n'ont pas réussi à déceler une relation entre les primes d'incitation et le nombre d'imputations pour les points de données manquants. Donc, bien que l'utilisation de primes d'incitation financière soit corrélée de manière significative avec la non-réponse totale (non-participation complète à l'enquête), nous ne dégageons aucune relation entre les primes d'incitation et la non-réponse partielle (réponse omise à certaines questions, mais non toutes). Donc, offrir une prime d'incitation à un répondant ne l'incite pas nécessairement à répondre plus diligemment à l'enquête.

L'analyse a révélé quelques résultats intéressants en ce qui concerne la relation entre certaines variables de contrôle et la non-réponse partielle. Alors que le niveau de scolarité, l'âge et le sexe ont été utilisés principalement comme



comment les aliments sont produits est relié au sujet plus général de l'enquête, nous pensons que le fait d'avoir rendu saillante la concentration de l'enquête sur l'agriculture et l'environnement dans notre demande de participation à l'enquête pourrait avoir suscité une plus grande diligence à remplir le questionnaire chez les répondants qui savaient, ou se souciaient suffisamment de savoir, comment les aliments sont produits. Cependant, notre analyse suggère aussi que le soutien pour le bien-être des animaux est positivement relié à la non-réponse partielle, ce qui est en contradiction avec l'hypothèse 3. Ces résultats font ressortir le besoin d'examiner de près les facteurs associés au sujet d'une enquête en tant que covariables possibles de la non-réponse partielle et de son corollaire, l'erreur due à la non-réponse.

Alors que l'enquête sur laquelle portait la présente étude se concentrait sur l'agriculture et l'environnement, nos résultats ayant trait au sujet de l'enquête pourraient avoir des incidences dans les enquêtes qui portent sur d'autres secteurs. Nous avons des raisons de croire que la non-réponse totale et partielle peut être influencée par la proximité par rapport à tout sujet d'enquête ou secteur d'activités étudié, ou par le niveau d'intérêt pour ce sujet ou secteur d'activités, surtout si cet aspect de l'enquête est rendu saillant durant la demande de participation. Par exemple, si une enquête se concentre sur l'industrie automobile et que cette caractéristique est rendue saillante durant la demande de participation, il est fort probable que cette information aura une incidence sur la courbe de réponse. Essentiellement, ces résultats donnent à penser que les chercheurs qui conçoivent les enquêtes doivent réfléchir de manière critique à la façon dont le contexte du sujet de l'enquête, tel que l'industrie ou le secteur sur lequel elle se concentre, pourrait influencer la participation de certaines sous-populations figurant dans la base de sondage. Bien que cette généralisation paraisse raisonnable, nous pensons que des études comparables axées sur d'autres secteurs seront nécessaires avant que l'on puisse tirer des conclusions catégoriques.

Nous discutons ensuite de la relation entre les primes d'incitation prépayées, d'une part, et la participation à l'enquête et la non-réponse partielle, d'autre part. En ce qui concerne la relation entre les primes d'incitation et la réponse, notre étude suggère que les primes d'incitation prépayées accroissent généralement la probabilité de participation à l'enquête, même si la proximité par rapport à l'agriculture et au paysage rural (contexte du sujet de l'enquête) est prise en compte. Nos résultats confirment l'hypothèse 4 et ceux publiés antérieurement (Singer et coll. 2000; Groves 2006; Church 1993; Trussell et Lavrakas 2004; Goyder 1982; Yu et Cooper 1983), car ils montrent que les répondants ayant reçu une prime d'incitation prépayée étaient significativement plus susceptibles de

Notre analyse révèle que la probabilité de participation à l'enquête sur l'agriculture et l'environnement examinée ici varie significativement selon la proximité des unités échantillonnées par rapport au paysage agricole et rural (lieu de résidence). Notre analyse corrobore notre première hypothèse et la proposition théorique du levier et de la saillance, car nous constatons que les résidents des cantons exurbains et des localités rurales sont tous significativement plus enclins à participer à l'enquête que ceux des localités urbaines. Le fait le plus susceptible d'expliquer le profil révélé par l'analyse est que les personnes résidant dans les cantons exurbains et les localités rurales ont plus de chances d'interagir avec le paysage agricole et rural que celles résidant dans les localités urbaines (voir tableau 2). Donc, selon nous, le taux de participation à l'enquête a été plus élevé chez les répondants résidant près du paysage agricole et rural à cause du levier positif du thème axé sur l'agriculture et l'environnement de l'enquête.

Nous observons également une certaine relation entre l'intérêt pour le sujet de l'enquête (mesuré par la proximité par rapport au paysage agricole et rural) et la qualité de la réponse (mesurée par la non-réponse partielle). À l'appui de notre seconde hypothèse, la présente étude apporte des preuves modestes que la non-réponse partielle varie selon la proximité par rapport au paysage agricole et rural. Dans le cas de la non-réponse partielle I, les données laissent entendre que les résidents des cantons exurbains sont moins susceptibles d'omettre de répondre à certaines questions que ceux des localités urbaines, tandis que les résidents des cantons exurbains et des localités rurales sont plus susceptibles d'omettre de répondre à certaines questions correspondantes. Les cas de non-réponse II. Les cas de non-réponse manquant associés aux questions exerçant la demande cognitive la plus grande (non-réponse partielle III) ne variaient pas selon le lieu de résidence (intérêt pour le sujet de l'enquête). Ces constatations donnent à penser que les résidents des localités plus rurales (cantons exurbains et localités rurales) produisent de moins bons résultats que ceux des localités urbaines quand les réponses manquant ont trait à des questions de l'enquête exerçant une demande cognitive de niveau moyen. Ce résultat est curieux, mais nous n'arrivons pas à expliquer pourquoi il en est ainsi. Une explication possible serait la différence de niveau de scolarité entre les résidents des localités urbaines et des localités rurales, mais l'étude inclut le contrôle statistique des effets du niveau de scolarité. D'autres travaux seront certainement nécessaires à cet égard.

Les connaissances sur la production des aliments, un autre indicateur de la proximité par rapport au paysage agricole et rural, sont liées négativement à la non-réponse partielle, ce qui concorde avec nos attentes (hypothèse 3) et la théorie du levier et de la saillance. Puisque savoir



Tableau 8  
Modèles de régression logistique<sup>a</sup> pour la non-réponse partielle

Non-réponse I <sup>b</sup>	Non-réponse II <sup>b</sup>	Non-réponse III <sup>c</sup>
Aucune manquante : logarithme du rapport des cotes	Aucune manquante : logarithme du rapport des cotes	Certaines manquantes : logarithme du rapport des cotes

Situation concernant la prime d'incitation		
N'a pas reçu de prime d'incitation	-	0,16
A reçu une prime d'incitation	(0,16)	0,10
<i>Intérêt du sujet – Lieu de résidence</i>		
Résidents des localités urbaines	-	0,54
Résidents des localités suburbaines	(0,26)	(0,29)
Résidents des villes/villages exurbains	(0,31)	0,30
Résidents des cantons exurbains	-0,74**	0,85**
Résidents des localités rurales	(0,27)	(0,30)
	(0,40)	0,82**
<i>Intérêt du sujet – Connaissances sur les aliments et bien-être des animaux</i>		
Connaissances sur la production des aliments	-0,07	-0,13*
Importance du bien-être des animaux	(0,05)	(0,05)
	0,10	0,09*
	(0,05)	(0,04)
<i>Variables de contrôle</i>		
Niveau de scolarité :		
Études secondaires ou moins	-0,79***	0,13
Études collégiales partielles	(0,20)	(0,19)
Baccalauréat	-1,08***	-0,32
	(0,23)	(0,27)
Études supérieures partielles/travail	-0,99***	0,12
	(0,24)	(0,24)
professionnel et plus élevé	0,03***	0,04***
Âge	(0,01)	(0,00)
Sexe (féminin = 1)	0,03	0,53**
	(0,17)	(0,17)
Race blanche	-0,38	-0,05
	(0,32)	(0,28)
<i>Statistiques du modèle</i>		
Ordonnée à l'origine	0,16	-3,56
	85,80	93,25
Khi-deux de Wald <sup>d</sup>		
N	828	828

Seuil de signification : \*\*\* < 0,001 ; \*\* < 0,01 ; \* < 0,05

<sup>a</sup> Nous avons testé l'existence d'effets d'interaction éventuels dans ces modèles entre le lieu de résidence et les primes d'incitation, entre l'âge et les primes d'incitation et entre l'ethnicité (blanche) et les primes d'incitation comme l'ont fait Singer et coll. (2000). Nous n'avons pu déclarer significatives aucune de ces interactions.  
<sup>b</sup> Les modèles de non-réponse partielle I et II sont des modèles logit à rapports des cotes proportionnels partiellement contraints, parce que certains prédicteurs de ces modèles violent l'hypothèse des droites parallèles. Nous avons donc permis à ces prédicteurs de varier, tandis que les autres ont été contraints. Nous avons utilisé le code de programmation du module gologit2 de stata de Williams (2006) pour estimer le modèle.  
<sup>c</sup> Ce modèle est un modèle de régression logistique dépendante binaire (variable recodée en deux catégories).  
<sup>d</sup> Les nombres de degrés de liberté sont 14, 14 et 13 pour les modèles à demande cognitive faible, moyenne et élevée, respectivement.

corrélées positivement à la non-réponse partielle (tableau 8, colonne 4). Comme le montre le tableau 8, une augmentation d'une unité de l'importance accordée au bien-être des animaux donne lieu à une augmentation de 0,09 unité du logarithme du rapport des cotes de la non-réponse partielle (surtout la non-réponse partielle II). Ce résultat ne concorde pas avec nos attentes.

En ce qui concerne les effets des primes d'incitation, nous ne dégageons aucune relation significative entre ces primes et n'importe laquelle des trois mesures de la non-réponse partielle (tableau 8, colonnes 2, 4 et 6), contrairement à nos attentes.

En ce qui concerne les variables de contrôle, nous constatons que le niveau de scolarité est relié significativement à la non-réponse partielle, ce qui corrobore les constatations antérieures de Singer et coll. (2000). Dans notre cas, les répondants qui ont fait des études collégiales partielles, titulaires d'un baccalauréat ou ayant fait des études supérieures partielles/un travail professionnel avait moins de chances (logarithme du rapport des cotes de -0,79, -1,08 et -0,99, respectivement) de ne pas répondre aux questions de l'enquête exerçant la demande cognitive la plus faible (non-réponse partielle I) que ceux n'ayant fait que des études secondaires complètes ou partielles (tableau 8, colonne 2).

Enfin, la non-réponse partielle liée aux questions de l'enquête dont la demande cognitive était comparativement plus élevée (c'est-à-dire la non-réponse partielle II et la non-réponse partielle III) ne variait pas selon le niveau de scolarité (tableau 8, colonnes 4 et 6). Nous dégageons également une relation positive entre l'âge et les trois mesures de la non-réponse partielle (tableau 8, colonnes 2, 4 et 6), ce qui concorde avec les résultats de Singer et coll. (2000).

Egalement en harmonie avec les travaux antérieurs de

Tableau 7  
Régression logistique<sup>a</sup> de la probabilité de participation

Logarithme du rapport des cotes de la participation		Erreur-type	
Situation concernant la prime d'incitation			
N'a pas reçu de prime d'incitation (réf.)	-	0,09	-
A reçu une prime d'incitation	0,73***		
Lieu de résidence			
Localité urbaine (réf.)	-		
Localité suburbaine	0,27*		0,13
Ville/village exurbain	0,25		0,15
Canton exurbain	0,60***		0,13
Localité rurale	0,37*		0,15
Première option pour répondre (sexe féminin = 1)	-0,05		0,09
Statistiques du modèle			
Ordonnée à l'origine	-1,42***		
$\chi$ de Wald (ddl = 6)	93,25***		

<sup>a</sup> Dans ce modèle, nous avons testé les effets d'interaction éventuels entre le lieu de résidence et les primes d'incitation. Nous n'avons pu déclarer significatives aucune de ces interactions.



### 6.3 Modèle de régression logistique pour la non-réponse partielle

Comme nous l'avons mentionné plus haut dans cette section, nous avons limité notre analyse de la non-réponse partielle à la modélisation multivariée et ce, principalement pour que l'article reste bref tout en respectant notre objectif d'évaluer les effets partiels de nos principales variables indépendantes. Les données laissent entendre que le levier attendu du sujet de l'enquête n'est que moyennement relié à la non-réponse partielle. En ce qui concerne la non-réponse partielle I (c'est-à-dire la variable créée d'après les questions exerçant la demande cognitive la plus faible sur les répondants à l'enquête), l'analyse suggère qu'elle est plus faible chez les répondants résidant dans les cantons exurbains (logarithme du rapport des cotes de -0,74) que chez ceux résidant dans les localités urbaines, bien que cette différence disparaisse aux valeurs plus élevées de cette variable (tableau 8, colonnes 2 et 3). Par contre, pour la non-réponse partielle II (la variable de non-réponse partielle créée d'après les questions les plus cognitivement exigeantes que celles utilisées pour créer la variable de réponse partielle I), nous constatons qu'elle est plus susceptible d'être élevée chez les résidents des cantons exurbains et des localités rurales (logarithmes des rapports des cotes de 0,85 et 0,82, respectivement) que chez les résidents des localités urbaines (tableau 8, colonne 4). En ce qui concerne la non-réponse partielle III (variable de non-réponse créée d'après les questions les plus cognitivement exigeantes), l'analyse ne révèle aucune différence significative selon le lieu de résidence, qui est notre mesure indirecte du niveau d'intérêt pour le sujet de l'enquête.

Confirmant l'effet prévu de l'intérêt pour le sujet de l'enquête sur la non-réponse partielle, l'analyse indique aussi que les connaissances des répondants sur la façon dont les aliments sont produits sont reliées significativement à la non-réponse partielle. Dans le cas de la non-réponse partielle II, les données montrent que le logarithme du rapport des cotes est plus faible (-0,13) pour les répondants qui ont déclaré savoir comment les aliments sont produits que pour ceux qui ont dit en savoir moins (tableau 8 colonne 4). Cette relation est plus forte aux valeurs plus élevées de la variable : les connaissances sur la production des aliments donnent un logarithme du rapport des cotes exprimant les chances de non-réponse partielle (-0,35) plus faible quand la valeur de la catégorie passe de 0 à 1 (tableau 8, colonne 5). Ce résultat suggère que le levier positif du sujet de l'enquête pourrait avoir pour effet une plus grande minutie dans la réponse au questionnaire chez les répondants ayant une meilleure connaissance de la façon dont les aliments sont produits. Nous constatons aussi que les opinions des répondants au sujet de l'importance du bien-être des animaux, un sous-thème essentiel de cette enquête particulière, sont

L'analyse de régression logistique semble également confirmer notre constatation antérieure voulant que la probabilité de participer varie significativement selon que l'unité échantillonnée a reçu ou non une prime d'incitation. Les chances de participer à l'enquête étaient plus élevées pour les répondants ayant reçu une prime d'incitation (logarithme du rapport des cotes de 0,73) que pour ceux qui n'en avaient pas reçu, en neutralisant les effets de la proximité par rapport à l'agriculture et au paysage agricole, ainsi que du sexe (féminin = 1) du membre du ménage sélectionné aléatoirement comme première option pour remplir et renvoyer le questionnaire (tableau 7).

Comme le statut socioéconomique varie significativement selon le lieu géographique de résidence (Lobao 1990) et une incidence sur la réponse à l'enquête (Davern et coll. 2003 ; Singer et coll. 2000), nous avons cherché à contrôler les effets éventuels du revenu par habitant et du revenu du ménage (statut socioéconomique) sur la probabilité de participer à l'enquête en utilisant un modèle linéaire hiérarchique (MLH). Pour cela, nous avons relié les répondants à leur groupe d'ilots et aux caractéristiques de ce dernier (spécialement, le revenu par habitant au niveau du groupe d'ilots et le revenu médian du ménage au niveau du groupe d'ilots) établies d'après les données du recensement de la population des États-Unis de 2000. Pour l'analyse MLH, nous avons d'abord estimé un modèle entièrement non conditionnel (c'est-à-dire une ANOVA) pour déterminer si la probabilité de participer à l'enquête variait de manière significative selon le groupe d'ilots. En modélisation linéaire hiérarchique, l'estimation d'un modèle entièrement non conditionnel (modèle sans les prédicteurs à tous les niveaux de l'analyse) est habituellement effectuée pour déterminer si la variable dépendante varie selon l'unité d'analyse de niveau 2 (ou plus élevée) telle que le quartier, le groupe d'ilots ou le district scolaire. Ce modèle initial (ANOVA) aide souvent les chercheurs à déterminer s'il faut procéder à l'analyse multivariée. Notre analyse MLH initiale (ANOVA) n'a révélé aucune variation significative de la probabilité de participer à l'enquête selon le groupe d'ilots ( $t_{au} = 0,04$  ;  $p = 0,493$ ). Bien que ce résultat donne à penser que la probabilité moyenne de participation à l'enquête est à peu près la même pour tous les groupes d'ilots malgré les différences de revenu par habitant et de revenu médian disponible du ménage, nous reconnaissons que ce modèle MLH pourrait être instable, étant donné que le nombre de cas échantillonnés par groupe d'ilots était généralement faible. C'est peut-être pour cette raison que nous ne constatons aucune variation significative de la probabilité de participation selon le groupe d'ilots (erreur de type II possible). Malgré ce problème éventuel de notre modèle entièrement non conditionnel, nous n'avons pas effectué d'analyse multivariée entièrement conditionnelle.



Notre analyse est en harmonie avec les résultats d'études antérieures, montrant aussi que les primes d'incitation prépayées accroissent significativement la participation à l'enquête (tableau 5). En dépit du fait que le contexte de l'enquête utilise pour notre analyse diffère sensiblement de celui des études antérieures examinant les effets des primes d'incitation, nous constatons que le taux de réponse pour les bénéficiaires d'une prime d'incitation contactés avec succès était de 43,7 % comparativement à 26,9 % pour les unités échantillonnées contactées avec succès qui n'ont pas reçu de prime d'incitation prépayée. Le test du khi-deux avec correction du second ordre de cette relation bivariable est également statistiquement significatif ( $\chi^2 = 73,8$ ;  $ddl = 1$ ;  $p = 0,000$ ). En fait, notre analyse donne à penser que l'élimination complète des primes d'incitation nuit au taux de participation dans toutes les catégories de répondants, quelle que soit la proximité par rapport au paysage agricole et rural, quoique cet effet soit le plus prononcé chez les résidents des localités urbaines (tableau 6). Cette constatation appuie notre pratique courante consistant à utiliser des primes d'incitation financière prépayées pour essayer ceux des localités urbaines.

L'analyse multivariée suggère en outre que la probabilité de participer à l'enquête varie significativement en fonction de la proximité par rapport à l'agriculture et au paysage rural, si l'on maintient statistiquement constants les effets de la situation concernant les primes d'incitation (a reçu ou n'a pas reçu de prime d'incitation). Les résidents des localités suburbaines, des cantons exurbains et des localités rurales sont significativement plus susceptibles de participer à l'enquête que ceux des localités urbaines (tableau 7). Par exemple, les résidents des cantons exurbains et des localités rurales ont plus de chances (logarithmiquement) de participer que ceux de 0,60 et de 0,37, respectivement) de participer que ceux des localités urbaines.

## 6.2 Modèle de régression logistique pour la participation à l'enquête

d'accroître les taux de réponse sans distinction entre les répondants vivant en région rurale ou en région urbaine. Elle confirme aussi l'importance des primes d'incitation en recherche par sondage.

Tableau 4  
Taux de participation selon le lieu de résidence

Lieu de résidence			
A répondu		N'a pas répondu	
Total <sup>a</sup>			
Localité urbaine	29,5 %	70,5 %	100 % (424)
Localité suburbaine	32,6 %	67,4 %	100 % (917)
Ville/village exurbain (constitué en corporation)	33,1 %	66,9 %	100 % (379)
Canton exurbain (non constitué en corporation)	40,5 %	59,5 %	100 % (684)
Localité rurale	35,8 %	64,2 %	100 % (405)
Total	35,4 %	65,6 %	100 % (2 809)

<sup>a</sup> Le nombre total de cas admissibles pour chaque catégorie de résidence est entre parenthèses.

Khi-deux avec correction du second ordre (3,7) = 14,2 ; P = 0,003 (correction pour les effets du plan de sondage)

Khi-deux avec correction du second ordre (3,7) = 14,2;  $P = 0,003$  (correction pour les effets du plan de sondage).  
<sup>a</sup> Le nombre total de cas admissibles pour chaque catégorie de résidence est entre parenthèses.

Tableau 5  
Réponse à l'enquête selon la situation concernant la prime d'incitation

Situation de prime d'incitation			
A répondu		N'a pas répondu	
Total <sup>a</sup>			
Prime d'incitation	43,7 %	56,3 %	100 % (1 410)
Pas de prime d'incitation	26,9 %	73,1 %	100 % (1 401)
Total	35,4 %	64,6 %	100 % (2 811)

<sup>a</sup> Le nombre total de cas admissibles selon la situation concernant la prime d'incitation est entre parenthèses.

Khi-deux avec correction du second ordre (1) = 73,8 ;  $P = 0,000$  (correction pour les effets du plan de sondage)

Khi-deux avec correction du second ordre (1) = 73,8;  $P = 0,000$  (correction pour les effets du plan de sondage).  
<sup>a</sup> Le nombre total de cas admissibles selon la situation concernant la prime d'incitation est entre parenthèses.

Tableau 6  
Taux de réponse selon la prime d'incitation et le lieu de résidence le long du continuum rural-urbain

A reçu une prime d'incitation			
N'a pas reçu de prime d'incitation		Différence entre les réponses	
Localité urbaine	0,41	0,19	0,22
Localité suburbaine	0,41	0,24	0,17
Ville/village exurbain	0,39	0,27	0,12
Canton exurbain	0,48	0,31	0,17
Localité rurale	0,44	0,27	0,17
Total	0,43	0,26	0,17

**Tableau 2**  
Fréquence des visites dans une ferme en exploitation

Lieu de résidence	Ohio Survey de 2006 <sup>a</sup>			Animal Welfare Survey de 2007 <sup>b</sup>		
	Jamais/ Occasionnellement/ Total <sup>c</sup>	fréquentement	Total <sup>c</sup>	Jamais/ Occasionnellement/ Total <sup>c</sup>	fréquentement	Total <sup>c</sup>
Localité urbaine	90,4 %	9,6 %	100 % (185)	19,0 %	100 % (121)	100 % (121)
Localité suburbaine	87,5 %	12,5 %	100 % (536)	16,3 %	100 % (285)	100 % (285)
Ville/village exurbain (constitué en corporation)	78,6 %	21,4 %	100 % (217)	76,4 %	100 % (124)	100 % (124)
Canton exurbain (non constitué en corporation)	74,9 %	25,1 %	100 % (434)	67,9 %	100 % (264)	100 % (264)
Localité rurale	73,1 %	26,9 %	100 % (238)	70,6 %	100 % (136)	100 % (136)
Total	80,6 %	19,4 %	100 % (1 610)	74,2 %	100 % (930)	100 % (930)

<sup>a</sup> Le nombre total de cas admissibles pour chaque catégorie résidentielle est entre parenthèses.  
<sup>b</sup> Khi-deux avec correction du second ordre (3,61) = 43,3 ;  $P = 0,0000$  (correction pour les effets du plan de sondage)  
<sup>c</sup> Khi-deux avec correction du second ordre (3,67) = 16,7 ;  $P = 0,001$  (correction pour les effets du plan de sondage)

**Tableau 3**  
Statistiques descriptives pour les variables de contrôle

Niveau de scolarité :		Moyenne/pour cent		Écart-type	
Études secondaires ou moins	Études collégiales partielles	36,8 %	-	-	-
	Baccalauréat	13,7 %	-	-	-
	Études postsecondaires/travail professionnel et plus élevé	17,2 %	-	-	-
	Sexe :				
Masculin	48,2 %	51,8 %	-	-	-
	Féminin	51,8 %	-	-	-
Ethnicité :	Blanche	91,7 %	8,3 %	-	-
	Non blanche	8,3 %	-	-	-
Âge :		51,9	49 842,3	15,8	25 258,7
Revenu moyen du ménage au niveau du groupe d'îlots, 2000		42 616,3	16 728,6		
Revenu médian du ménage au niveau du groupe d'îlots, 2000		42 616,3	16 728,6		

## 6. Résultats

### 6.1 Résultats bivariés pour la participation à l'enquête

L'analyse bivariable donne à penser que la participation à l'enquête varie de manière significative en fonction de la proximité par rapport au paysage agricole et rural (lieu de résidence le long du continuum rural-urbain). Comme le montre le tableau 4, les répondants résidant dans des localités géographiquement plus rurales (résidents des cantons ruraux et exurbains) ont un taux de participation à l'enquête plus élevé que ceux résidant dans des localités géographiques plus urbaines (résidents des localités exurbaines et des localités suburbaines). L'analyse montre aussi que ceux résidant dans les localités constituées en corporations exurbaines intermédiaires (villes et villages) étaient un peu plus susceptibles de participer à l'enquête que les résidents des localités urbaines. Un test du khi-deux avec correction du second ordre (Rao et Scott 1984) de la relation entre la participation à l'enquête et le lieu de résidence a produit un résultat significatif ( $\chi^2 = 14,2$  ; ddl = 3,7 ;  $p = 0,003$ ).

À fin d'évaluer la participation à l'enquête, nous avons recouru à l'analyse bivariable (tableaux de contingence) et à la modélisation de régressions logistiques. Pour les tableaux de contingence, nous avons utilisé comme statistique le khi-deux de Pearson corrigé des effets du plan de sondage au moyen de la correction du deuxième ordre de Rao et Scott (1984). Nous utilisons cette correction parce que les caractéristiques du plan de sondage, comme la stratification et la mise en grappes, peuvent avoir une influence sur les tests d'association (Lohr 1999). Afin de limiter la longueur de l'article, nous suivons un plan d'analyse différent pour l'ensemble de variables de non-réponse partielle. Pour cet ensemble, nous effectuons uniquement une analyse multivariée (régression logistique). Le passage direct à l'analyse multivariée nous permet d'examiner les effets partiels des divers prédicteurs utilisés dans les modèles tout en maintenant la brièveté de l'article.



croissance monotone en ce qui concerne la proximité par rapport à l'agriculture et au paysage rural, mais à l'intérieur des blocs le profil est moins certain. De nouveau, ici, nous ne fournissons pas de statistique descriptive pour cette variable, car la section réservée à l'analyse donne une très bonne idée de la façon dont elle est distribuée.

#### *Connaissance de la production des aliments et soutien pour le bien-être des animaux : Deux autres indicateurs du levier exercé par le sujet de l'enquête utilisés dans l'analyse*

sont les deux questions destinées à mesurer les connaissances des unités échantillonnées concernant la production des aliments qu'elles consomment et leurs opinions au sujet de l'importance du bien-être des animaux. La première question était « Dans quelle mesure savez-vous comment vos aliments sont produits ? Veuillez indiquer sur une échelle allant de 1 à 7 votre niveau de connaissance. » Pour cette question, le score moyen était de 4,47 avec un écart-type de 1,60. La deuxième question était : « En pensant aux animaux d'élevage en général, quelle importance cette question a-t-elle pour vous ? Veuillez l'indiquer sur une échelle allant de 1 (pas importante) à 7 (très importante). » Pour cette question, le score moyen était de 4,50 avec un écart-type de 1,68. Ces deux indicateurs ne sont utilisés que dans les analyses portant sur les variables de non-réponse partielle.

*Situation concernant les primes d'incitation : La situation des unités échantillonnées concernant les primes d'incitation (a reçu ou n'a pas reçu une prime d'incitation) est une variable indépendante principale dans les modèles de régression. La situation concernant les primes d'incitation est codée comme une variable binaire prenant les valeurs 0 (n'a pas reçu de prime d'incitation) et 1 (a reçu une prime d'incitation). Une fois de plus, nous ne fournissons aucune statistique descriptive pour cette variable, parce que l'analyse donne une bonne idée de sa distribution.*

*Variables de contrôle : Les variables de contrôle opérationnalisées dans une ou plusieurs des analyses effectuées dans la présente étude comprennent l'âge (l'âge du répondant à son dernier anniversaire) ; le niveau de scolarité (plus haut niveau de scolarité atteint), l'ethnicité (race blanche = 1 ; toutes les autres = 0) et le sexe (masculin = 0 et féminin = 1), ainsi que le revenu par habitant et le revenu médian disponible du ménage au niveau du groupe d'ilots auquel appartenaient chaque unité échantillonnée, d'après les données du recensement de la population de 2000. Nous introduisons des variables de contrôle pour les effets de ces variables, parce que des études antérieures indiquent qu'elles peuvent avoir une incidence sur la non-réponse partielle (Davern et coll. 2003 ; Singer et coll. 2000). Les statistiques descriptives pour ces variables purement de contrôle sont présentées au tableau 3.*

de notre enquête agisse comme un levier positif sur les unités échantillonnées résidant près des zones agricoles et du paysage rural. Même s'il ne s'agit pas d'une mesure directe du levier, elle est en harmonie avec la suggestion de Groves et coll. (2000) selon laquelle l'effet de levier qu'exerce une enquête donnée sur une unité échantillonnée peut être mesuré indirectement en s'appuyant sur des caractéristiques pertinentes de cette unité. En utilisant les caractéristiques spatiales de résidence des unités échantillonnées, nous nous appuyons sur le fait que celles résidant dans des régions plus rurales et en pleine campagne sont plus susceptibles d'avoir une interaction sociale et physique avec le paysage agricole et rural que celles résidant dans des lieux plus urbanisés (voir le tableau 2). En 2006 ainsi qu'en 2007, de plus fortes proportions de résidents des cantons exurbains et des régions rurales (une combinaison de villes/villages ruraux et de cantons ruraux) que de résidents des localités urbaines ont visité une ferme en exploitation, comme le montre le tableau 2. Nous reconnaissons que cela peut poser problème d'utiliser l'information provenant de nos propres répondants pour montrer l'association entre le lieu de résidence et les visites dans des fermes. Cependant, cette information est corroborée par celle provenant d'un échantillon différent, celui de la Ohio Survey de 2006.

Afin de déterminer le lieu de résidence des unités échantillonnées, nous avons géocodé le lieu de résidence de chaque répondant et l'avons assigné à l'un de quatre secteurs de localisation – urbain, suburbain, exurbain ou rural – en utilisant le géocodage ArcView de l'ESRI. Les unités échantillonnées appartenant aux secteurs exurbain et rural ont été en outre classifiées comme résidant dans des localités constituées en corporation (ville/village) ou dans des localités cantonales (plaine campagne). Ce mode de classification des unités échantillonnées comme vivant dans des localités urbaines, suburbaines, exurbaines ou rurales a été employé antérieurement avec succès dans le domaine de la science régionale (Audrac 1999 ; Sharp et Clark 2008).

Dans la présente étude, cette variable a été groupée en cinq catégories : 1) localité urbaine, 2) localité suburbaine, 3) ville/village exurbain, 4) canton exurbain et 5) localité rurale (villes/villages et cantons). Le classement par ordre des catégories ne suggère pas l'existence d'un ordre croissant monotone en fonction de la proximité par rapport à l'agriculture et au paysage rural entre les catégories 1 à 5. Cette variable devrait plutôt être considérée comme une variable nominale dont les catégories peuvent être groupées en blocs en se basant sur la proximité par rapport à l'agriculture et au paysage rural : bloc 1 (catégories 1 et 2) ayant la proximité la plus faible, bloc 2 (catégorie 3) ayant une proximité intermédiaire et bloc 3 (catégories 4 et 5) ayant la plus grande proximité. D'un bloc à l'autre, les catégories présentent une



de classification dans la catégorie « nombreuses réponses manquantes » (tableau I). Afin de vérifier si notre regroupement de ces variables masquait des variances dans la non-réponse partielle à l'intérieur des groupes (cas regroupés) pouvant être expliquées par nos deux variables indépendantes (lieu de résidence, c'est-à-dire un indicateur de l'intérêt pour le sujet de l'enquête, et prime d'incitation), nous avons effectué une analyse de variance à un facteur de ces cas groupés. À l'intérieur de ces groupes, aucune des trois variables de non-réponse partielle ne variait de manière significative selon le lieu de résidence ou l'offre d'une prime d'incitation. Les statistiques descriptives pour les trois variables de non-réponse partielle sont présentées au tableau I.

Tableau I

Statistiques descriptives des variables de non-réponse partielle			Non-réponse			Non-réponse		
			partielle I			partielle II		
			partielle III					
			Statistiques avant le recodage			Statistiques après le recodage		
			par groupe					
			Aucune manquante			Aucune manquante		
			Quelques-unes manquantes			Quelques-unes manquantes		
			Nombreuses manquantes			Nombreuses manquantes		
N	971	971	971	971	971	971	971	971
Moyenne	3,11	2,34	3,11	2,34	3,11	2,34	2,34	1,60
Ecart-type	5,06	5,93	5,06	5,93	5,06	5,93	5,93	3,25
Valeur minimale	0	0	0	0	0	0	0	0
Valeur maximale	44	48	44	48	44	48	48	29
Statistiques avant le recodage	30,07 %	59,53 %	7,62 %	32,65 %	7,83 %	43,36 %	54,69 %	1,96 %

5.3 Opérationnalisation des variables indépendantes et de contrôle

*Lieu de résidence* : Le thème de l'enquête axé sur les questions agricoles et environnementales a été rendu saillant durant la demande de participation à l'enquête (au moyen de la lettre de préavis, des lettres de couverture et de la conception du questionnaire), ce qui peut avoir une incidence négative ou positive sur la participation, selon le lieu de résidence du répondant le long du continuum rural-urbain. Le lieu de résidence est un indicateur de la proximité sociale et physique différentielle du répondant par rapport à l'agriculture et au paysage rural, parce que la proximité peut accroître les interactions sociales et (ou) physiques avec le sujet de l'enquête. L'association entre la proximité et la préoccupation au sujet de l'environnement a été proposée et testée à de nombreuses reprises par les spécialistes des sciences sociales (Dunlap et Hefterman 1975 ; Freudenbourg 1991 ; Sharp et Adua 2009). Nous faisons un pas de plus et émettons l'hypothèse de différences attitudeles associées à la proximité, et prévoyons des niveaux différents de participation à l'enquête ; en effet, nous supposons que les unités échantillonnées résidant plus près des régions agricoles et du paysage rural auront un taux élevé de participation à l'enquête que celles vivant dans les localités urbaines. Par conséquent, nous nous attendons à ce que le sujet spécialisé

Davern et coll. 2003 ; Kalderberg, Koenig et Becker 1994). Pour calculer la non-réponse partielle, nous avons calculé le nombre total de points de données manquants pour l'ensemble des répondants participant à l'enquête sur trois sous-ensembles de questions figurant dans le questionnaire afin de produire trois variables de non-réponse partielle : non-réponse partielle I, non-réponse partielle II et non-réponse partielle III. La variable de non-réponse partielle I a été créée pour les questions qui, à notre avis, exerçaient sur les répondants la demande cognitive comparativement la plus faible, y compris des questions telles que celles sur les caractéristiques démographiques et les questions d'opinion ne nécessitant pas beaucoup d'inspection. La variable de non-réponse partielle II a été créée pour les questions qui exerçaient sur les répondants une demande cognitive comparative plus élevée que celles utilisées pour créer la variable de non-réponse partielle I, comme celles nécessitant des efforts de remémoration importants et les questions d'opinion demandant un haut niveau d'inspection. La troisième variable est construite en s'appuyant sur les questions exerçant sur les répondants la demande cognitive comparative la plus élevée, comme des questions sur les connaissances et des questions nécessitant une certaine compréhension des concepts associés à l'élevage.

Dans les sommations sur ces variables, nous n'avons pas traité les réponses « Ne sait pas » comme une non-réponse partielle, étant donné que le questionnaire contenait une ou deux questions sur les connaissances pour lesquelles une réponse « Ne sait pas » pouvait être légitime. La variable de non-réponse partielle ne contient pas non plus les cas de « refus de répondre », car cette option n'a pas été offerte dans les questions utilisées pour créer les variables. Nous avons également exclu de ces variables les questions que les répondants avaient l'instruction d'omettre s'ils jugeaient qu'elles n'étaient pas applicables. Étant donné que la distribution de ces variables était fortement asymétrique (voir le tableau I), les variables de non-réponse partielle I et de non-réponse partielle II ont été regroupées en trois catégories ordinales (0 = aucune réponse manquante ; 1 = certaines réponses manquantes ; 2 = nombreuses questions manquantes) et analysées en utilisant le logit ordonné généralisé. La première catégorie (0) englobait les cas sans aucune non-réponse partielle, tandis que la deuxième catégorie (1) englobait les cas comptant de 1 à 9 occurrences de non-réponse. La troisième catégorie (2) comprenait les cas comptant 10 occurrences ou plus de non-réponse. Pour les besoins de notre analyse, nous avons également regroupé la variable de non-réponse partielle III en deux catégories : 0 (aucune réponse manquante) et 1 (une ou plusieurs réponses manquantes). Cette variable a été regroupée différemment des deux premières, parce que très peu de cas (19 seulement) satisfaisaient les critères

auprès du public de l'Ohio. Tout comme Groves et coll. (2000) s'attendaient à observer un effet de l'engagement communautaire sur les niveaux de participation, nous attendions à ce que les ménages résidant dans un lieu très proche de l'agriculture et du paysage rural aient un taux élevé de participation dans notre étude, indépendamment de la prime d'incitation, peut être au point qu'une prime d'incitation financière symbolique pourrait être jugée inutile pour les futures éditions de l'enquête.

5.1 Stratégie d'analyse

Deux ensembles d'analyses statistiques sont effectués dans la présente étude. Le premier ensemble est axé sur la participation à l'enquête (taux de réponse). Pour commencer, nous examinons la proportion d'unités échantillonnées contactées avec succès qui ont rempli et renvoyé le questionnaire, selon le lieu de résidence le long du continuum rural-urbain, qui est une mesure indirecte de la proximité géographique par rapport aux régions agricoles et rurales de l'État (hypothèse que nous justifions dans une section ultérieure) et selon la situation concernant la prime d'incitation. Conformément aux lignes directrices de l'American Association of Public Opinion Research (AAPOR) de 2008 pour les codes de classification, nous définissons les unités échantillonnées contactées avec succès comme étant i) celles pour lesquelles nous avons reçu un questionnaire rempli à la fin de la phase de collecte des données du projet et ii) celles pour lesquelles nous n'avions reçu en retour ni un questionnaire rempli ni la trousse d'enquête accompagnée d'un avis de colis non distribuable du United States Postal Service (USPS). Lors de nos communications avec l'USPS, nous avons demandé que tous les envois qui ne pouvaient être distribués à cause d'une adresse incorrecte ou de l'absence d'une adresse de réexpédition nous soient renvoyés. Les unités échantillonnées auxquelles ces envois non distribuables étaient adressés ont été considérées comme les unités avec lesquelles nous n'avons pas réussi à prendre contact. Nous recourons aussi à la régression logistique pour analyser plus en détail la probabilité de participation à l'enquête (code 1 = a répondu ; 0 = n'a pas répondu), en utilisant le lieu de résidence le long du continuum rural-urbain et la situation concernant la prime d'incitation comme prédicteurs principaux et en neutralisant simultanément les effets du statut socioéconomique au niveau du groupe d'îlots du répondant défini conformément au recensement de la population des États-Unis de 2000. Nous neutralisons l'effet du statut socioéconomique, parce que des études antérieures laissent entendre qu'il existe une certaine relation entre celui-ci et la participation aux enquêtes (Davern et coll. 2003 ; Singer et coll. 2000).

Le deuxième ensemble d'analyses porte sur la non-réponse partielle. Nous avons procédé à une analyse par

régression logistique ordonnée proportionnelle partielle (logit ordonné généralisé) sur les deux premières variables de non-réponse partielle (0 = aucune question manquante ; 1 = quelques questions manquantes ; 2 = de nombreuses questions manquantes), en employant de nouveau le lieu de résidence le long du continuum rural-urbain et la situation concernant la prime d'incitation comme variables indépendantes principales tout en neutralisant les effets de plusieurs autres variables. Nous utilisons le logit ordonné généralisé (rapports des cotes partiellement proportionnels) plutôt que le logit ordonné, parce que certains prédicteurs figurant dans ces modèles violaient l'hypothèse des rapports des cotes proportionnels de la régression logistique ordonnée. En utilisant un modèle à rapports des cotes partiellement proportionnels, nous arrivons à contraindre la relation entre les variables indépendantes et dépendantes satisfaisant à l'hypothèse des rapports des cotes proportionnels de la régression logistique ordonnée, tout en permettant aux relations ne satisfaisant pas à cette hypothèse de varier. Pour analyser la troisième variable de non-réponse, nous avons employé une régression logistique. Cette variable a été recodée sous forme d'une variable binaire (voir la section sur l'opérationnalisation des variables pour plus de renseignements).

5.2 Opérationnalisation des variables dépendantes

*Participation à l'enquête* : La participation à l'enquête (taux de réponse) est mesurée en calculant le nombre de questionnaires remplis renvoyés par les répondants (cas admissibles participants) en proportion du nombre d'unités échantillonnées contactées avec succès (total des cas admissibles). Cette mesure de la participation à l'enquête est conforme aux lignes directrices de l'AAPOR pour la mesure des taux de réponse. Les questionnaires non livrables retournés par l'USPS sans renseignements supplémentaires tels qu'une adresse de réexpédition ou une correction d'adresse ont été traités comme des cas inadmissibles. Les cas pour lesquels nous avons reçu le questionnaire rempli soit toute nouvelle information au sujet du cas fournie par l'USPS ont été traités comme des cas admissibles sur la base des définitions normalisées révisées de 2008 des codes de classement des cas et des taux de résultat recommandés par l'AAPOR. Pour exécuter l'analyse par régression logistique de la probabilité de réponse, nous avons donné à toutes les unités échantillonnées contactées avec succès (cas admissibles) le code de 1 (a retourné un questionnaire rempli) ou de 0 (n'a pas retourné un questionnaire rempli). Nous ne fournissons aucune statistique descriptive pour cette variable ici, car la section réservée à l'analyse, spécialement les valeurs de marge des tables de contingence, donne une bonne idée de la distribution de cette variable. *Qualité de la réponse* : La qualité de la réponse est mesurée par l'occurrence de la non-réponse partielle (voir



voie du biais d'optimisme ou de pessimisme. Étant donné ces mises en garde et le fait que la plupart des travaux antérieurs portant sur la relation entre les primes d'incitation et la participation aux enquêtes étaient fondés sur une analyse bivariable (prime d'incitation et participation à l'enquête), nous jugeons nécessaire de réexaminer l'effet des primes d'incitation sur la non-réponse aux enquêtes en tenant compte du lieu de résidence dans le contexte spatial et socioéconomique. Donc, en nous inspirant de ces travaux de recherche sur la façon dont les primes d'incitation sont liées à la participation aux enquêtes et à la non-réponse partielle, nous formulons les hypothèses qui suivent.

1. Les répondants qui reçoivent une prime d'incitation auront un taux de participation à l'enquête plus élevé que ceux qui n'en reçoivent pas, en tenant compte des effets de la proximité par rapport au paysage agricole et rural, et du statut socioéconomique.

2. Les primes d'incitation seront négativement associées à la non-réponse partielle. Autrement dit, les questionnaires auxquels auront répondu les personnes ayant reçu une prime d'incitation contiendront un plus petit nombre de points de données manquants que ceux remplis par des personnes qui n'ont pas reçu de prime d'incitation, en neutralisant les effets de la proximité du répondant par rapport au sujet de l'enquête et d'autres covariables.

## 5. Conception de l'étude

Le présent article s'appuie sur un sondage d'opinion publique portant sur des questions relatives aux aliments, à l'agriculture et à l'environnement, en se concentrant spécialement sur le bien-être des animaux d'élevage. L'enquête avait pour population cible les ménages de l'Ohio. Un échantillon initial de 3 000 répondants (avec leur adresse de résidence) a été sélectionné par échantillonnage aléatoire stratifié, la moitié (1 500) provenant des 22 comtés métropolitains à noyau urbain, et la deuxième moitié (1 500), des 66 comtés métropolitains périphériques et non métropolitains. Le nombre de ménages dans les comtés métropolitains à noyau urbain diffèrait de ceux dans les comtés métropolitains périphériques et non métropolitains, donnant donc un échantillon aléatoire disproportionné. Afin de tenir compte des probabilités inégales de sélection dans les deux strates, nous avons effectué une analyse pondérée pour les besoins du présent article.

Nous avons obtenu l'échantillon utilisé pour l'étude auprès de la firme Experian, un bureau d'évaluation du crédit et vendeur de listes privées ayant son siège aux États-Unis. L'échantillon a été tiré d'une base de sondage (base de données) comprenant les ménages de l'Ohio ainsi que leurs

adresses de résidence. Nous ne prétendons, certes, pas que cette base de sondage couvre tous les ménages de l'Ohio, mais nous pensons qu'il s'agit de l'une des listes et bases de données les plus fiables et les plus à jour à partir desquelles on peut tirer un échantillon aux États-Unis. Selon Experian, la base de données est mise à jour mensuellement.

L'enquête a été réalisée suivant un plan personnalisé modifié (Dillman 2000) comprenant jusqu'à quatre envois par la poste aux répondants prospectifs durant le printemps de 2007. Le premier envoi était une lettre de préavis envoyée à chaque unité échantillonnée, qui a été suivie peu à près par l'envoi d'une trousse d'enquête. Le troisième envoi était une carte de rappel adressée aux répondants pour les remercier de leur participation à l'étude ou les encourager à remplir le questionnaire et à le renvoyer, s'ils ne l'avaient pas encore fait. Le quatrième envoi, contenant une trousse d'enquête de remplacement, a été adressé aux répondants qui n'avaient pas renvoyé le questionnaire dûment rempli dix jours environ après l'envoi de la carte de rappel. Trois de ces quatre prises de contact avec les répondants contenaient des renseignements axés spécialement sur le sujet ou le thème de l'enquête. La lettre de préavis et les lettres de couverture accompagnant la trousse d'enquête initiale et celle de remplacement communiquaient spécifiquement aux répondants le sujet de l'enquête. En outre, le graphisme figurant sur la page de couverture du questionnaire (image d'animaux d'élevage) avait été sélectionné afin de renforcer la communication de ce sujet.

Les adresses des unités échantillonnées ont été géocodées et classées dans un secteur local (voir les détails plus loin dans cette section) afin de les localiser géographiquement sur le continuum rural-urbain. Cela nous a permis d'effectuer une analyse pour déterminer la relation existante entre la proximité des unités échantillonnées par rapport au paysage agricole et la probabilité de participer à l'enquête. Nous reconnaissons que certains citadins peuvent avoir de fréquentes interactions sociales et physiques avec l'agriculture et le paysage rural; cependant, ce genre d'interaction et ses effets sur le soutien de l'agriculture et de l'environnement sont les plus prononcés chez les personnes résidant dans des lieux plus ruraux et en pleine campagne (Freudenbourg 1991; Sharp et Adua 2009). Nous avons intégré dans l'enquête une expérience randomisée comportant des primes d'incitation. La première trousse d'enquête, envoyée à une moitié sélectionnée aléatoirement des unités échantillonnées, contenait une prime d'incitation de 2 \$ (deux billets d'un dollar), tandis que l'autre moitié de l'échantillon a reçu la même trousse, mais sans prime d'incitation. L'objectif pragmatique de cette expérience était d'évaluer l'efficacité de notre pratique consistant à inclure de modestes primes d'incitation financière dans les trousse d'enquête en vue d'améliorer la participation à nos enquêtes permanentes



que le domaine spécialisé sur lequel porte l'enquête que nous étudions diffère de celui de nombreuses études antérieures, nous jugeons nécessaire d'évaluer les effets des primes d'incitation sur la participation à l'enquête en même temps que nous examinons la relation entre la proximité par rapport à l'agriculture (le thème contextuel de notre enquête) et la réponse. En outre, nous croyons qu'il est important d'évaluer périodiquement l'utilité de l'utilisation de primes d'incitation dans la recherche par sondage, même si beaucoup d'attention a été accordée à ce sujet dans le passé.

Une autre question importante liée aux primes d'incitation est l'effet négatif possiblement plus prononcé sur la non-réponse partielle qui peut avoir lieu lorsqu'on incite des répondants réticents à participer à une enquête (voir Hansen 1980). Le dommage potentiel tient au fait que recourir à des moyens de persuasion tels que ces primes pourrait mener à recueillir de l'information auprès de répondants qui sont négligents ou indifférents quand ils répondent aux questions, nuisant ainsi en fin de compte à la qualité de l'information obtenue de cette façon (Singer et coll. 2000). En réponse à cette préoccupation, un certain nombre d'études ont pour but d'examiner la relation entre les primes d'incitation et la non-réponse partielle, et bon nombre d'entre elles donnent à penser que les primes d'incitation ne nuisent pas sérieusement à la qualité de la réponse; autrement dit, les primes d'incitation ne produisent pas des taux plus élevés de non-réponse partielle (Singer et coll. 2000; Singer et coll. 1998; Shettle et Mooney 1999; Davern et coll. 2003). En fait, Singer et coll. (2000) rapportent que les primes d'incitation prépayées aident à réduire la non-réponse partielle, qui est une mesure indirecte souvent utilisée de la qualité de la réponse ou des données. Cependant, ils signalent aussi que les répondants qui ont reçu des primes d'incitation étaient plus susceptibles de donner des réponses optimistes dans certains cas et d'être plus pessimistes dans d'autres (au sujet de variables différentes). En ce qui concerne notre enquête, une préoccupation essentielle est que les citadins qui sont incités à participer pourraient fournir des données de moins bonne qualité (mesurée par la non-réponse) que les répondants plus proches du paysage agricole et rural.

Pour résumer la revue, nous constatons que les travaux de recherche donnent généralement à penser que les primes d'incitation aident à améliorer les taux de réponse aux enquêtes, en n'ayant que peu d'effet, voire aucun, sur la non-réponse partielle. Bien qu'il en soit habituellement ainsi, certains résultats s'écartent de cette attente (Church 1993). En outre, alors que les auteurs de nombreuses études constatent que l'offre de primes d'incitation prépayées n'a pas d'incidence sur la non-réponse partielle, les travaux de Singer et coll. (2000) suggèrent que donner des primes d'incitation peut compromettre la qualité des données par la

d'incitation et des lettres de rappel) peuvent susciter une réciprocité de la part des ménages échantillonnés en termes de propension à participer à une enquête. En outre, Weisberg (2005) constate que l'échange social est une théorie qui pourrait expliquer la relation entre les primes d'incitation et la participation aux enquêtes, observant que, dans cette perspective, donner aux répondants une prime d'incitation financière pour qu'ils participent à l'enquête peut être perçu comme une gentillesse qui évoque une norme de réciprocité (page 165).

Afin de concevoir « des façons et des moyens » d'accroître les taux de réponse aux enquêtes et pour éprouver la théorie des échanges sociaux en regard de l'utilisation de primes d'incitation dans la recherche par sondage, un certain nombre d'études expérimentales ont été réalisées en vue d'examiner la relation entre l'offre de primes d'incitation aux répondants et la participation aux enquêtes. Alors que certaines de ces études portaient principalement sur les effets des primes d'incitation sur le taux de réponse et la non-réponse partielle (Groves, Couper, Presser, Singer, Tourangeau, Acosta et Nelson 2006; Trussell et Lavrakas 2004; James et Bolstein 1992; Church 1993; Singer 2000; Yammarino, Skinner et Childers 1991; Fox, Crask et Kim 1988), d'autres examinaient les effets des primes d'incitation sur les attentes des répondants et leurs opinions au sujet des enquêtes (James et Bolstein 1990; Singer et coll. 1998). Corroborant la proposition principale de la théorie des échanges et la norme de réciprocité, nombre de ces études révèlent une relation positive entre les primes d'incitation et les taux de réponses (Singer et coll. 2000; Groves, Couper, Presser, Singer, Tourangeau, Acosta et Nelson 2006; Church 2006; Trussell et Lavrakas 1993; Trussell et Lavrakas 2004; Goyder 1982; Yu et Cooper 1983).

Bien que de nombreuses études confirment l'importance des primes d'incitation pour ce qui est d'encourager la participation aux enquêtes, le verdict fondé sur les données empiriques quant à la relation entre les primes d'incitation et la participation aux enquêtes est loin d'être unanime. Dans une méta-analyse des études expérimentales et quasi expérimentales portant sur diverses mesures d'incitation, Church (1993) mentionne que 1 % des études couvertes par l'analyse n'avaient mis en évidence aucune preuve que les primes d'incitation ont une incidence sur la participation. Il signale aussi que 10 % des 74 études analysées ont effectivement révélé une relation négative entre les mesures d'incitation et la participation à l'enquête. En fait, c'est en partie face à cette réalité que Groves et coll. (2000) ont proposé la théorie du levier et de la saillance pour essayer d'expliquer pourquoi « les primes d'incitation fonctionnent parfois » et « parfois pas » (page 299). Étant donné que les constatations concernant les effets des primes d'incitation sur la participation aux enquêtes sont quelque peu contradictoires, et

D'autre part, la probabilité qu'une unité échantillonnée accepte de participer, toutes choses étant égales par ailleurs, perçoit négativement sont rendus saillants durant la demande de participation à l'enquête.

Groves et coll. (2000) appuient empiriquement cette position théorique. Ils présentent l'engagement civique (mesuré par la participation à la vie communautaire) et les mesures d'incitation comme des leviers agissant sur la participation à l'enquête et réussissent à montrer que ces deux attributs influent positivement sur la probabilité de participation, l'effet des primes d'incitation diminuant chez les unités échantillonnées dont l'engagement civique est plus important. En utilisant l'engagement civique comme mesure de l'effet de levier d'une enquête sur les unités échantillonnées, Groves et coll. (2000) reconnaissent que l'effet de levier n'est pas mesuré directement. Il peut plutôt être glané en considérant certaines caractéristiques des répondants relatives à l'enquête ou les caractéristiques de celle-ci, qui peuvent exercer une influence positive ou négative sur la probabilité de participation. Il y a aussi des indications que si les demandes de participation à l'enquête sont personnalisées en fonction des préoccupations des unités échantillonnées ou de ce qu'elles considèrent comme important, la probabilité de leur participation est accrue (Dillman 2000 ; Groves et Couper 1998).

En nous appuyant sur la proposition théorique du levier de la saillance, nous nous attendons à observer des taux de participation plus élevés chez les répondants dont les caractéristiques les rendent plus susceptibles de percevoir positivement des attributs. De manière correspondante, nous nous attendons aussi à ce que les unités dont les caractéristiques les rendent moins susceptibles de percevoir positivement ces attributs aient un taux plus faible de participation à l'enquête. Dans notre domaine particulier de recherche, nous prévoyons que la mesure dans laquelle les unités échantillonnées sont proches du paysage agricole et rural (le thème contextuel de notre enquête permanente) aura une incidence sur la participation à l'enquête et sur la non-réponse partielle. Ce raisonnement s'applique également à nos attentes au sujet des répondants qui font état d'une connaissance plus vaste de la façon dont les aliments sont produits et qui accordent également de l'importance au bien-être des animaux (un sous-thème essentiel de cette étude particulière). Nous nous inspirons donc de la théorie du levier et de la saillance pour proposer les hypothèses qui suivent.

1. La concentration de notre enquête sur l'agriculture et l'environnement, qui a été rendue saillante dans la conception de l'enquête, devrait avoir un effet de levier positif sur les répondants qui sont plus proches, socialement et physiquement, de l'agriculture et de

l'environnement rural (c'est-à-dire ceux qui résident dans des localités plus rurales). Nous émettons donc l'hypothèse que le taux de participation variera en fonction du lieu de résidence.

2. Nous nous attendons à ce que les répondants qui sont plus proches de l'agriculture et du paysage rural répondent à l'enquête de manière plus appliquée que ceux qui n'en sont pas proches, car les premiers sont plus susceptibles d'être motivés par le sujet de l'enquête (c'est-à-dire sont effet de levier positif). Nous émettons donc l'hypothèse que la non-réponse partielle variera selon la proximité par rapport à l'agriculture et au paysage rural.

3. Les unités échantillonnées qui ont une plus vaste connaissance de la façon dont les aliments sont produits, ainsi que celles qui estiment que le bien-être des animaux est important produiront un moins grand nombre de non-réponses partielles. Ces répondants manifesteront sans doute un plus grand intérêt pour les thèmes de l'agriculture et de l'environnement sur lesquels porte l'enquête et, par conséquent, répondront à l'enquête avec plus de diligence.

#### 4. Primes d'incitation et participation à l'enquête

L'utilisation de diverses formes de prime d'incitation, particulièrement des primes d'incitation (financières) prépayées est devenue une pratique courante de la recherche par sondage. Alors que la raison pratique de l'offre de primes d'incitation aux unités échantillonnées est de les encourager à participer, l'origine théorique de cette pratique remonte, en partie, à la théorie des échanges sociaux (Dillman 1978). Cette théorie repose sur le postulat que les actions d'une personne soient motivées principalement par le rendement qu'elle en attend ou qu'elle en obtient (Weisberg 2005). Gouldner (1960) se penche sur la norme de réciprocité, qui s'apparente à la théorie des échanges sociaux, observant que, dans la mesure où les hommes se conforment à une telle règle de réciprocité, quand une partie agit d'une manière qui profite à une autre, une obligation est créée. Le bénéficiaire est alors *endetté* à l'égard du donneur et le demeure jusqu'à ce qu'il rembourse sa dette (page 174). Dans la perspective de Gouldner, la norme de réciprocité exerce deux demandes sur les individus : 1) les individus devraient aider ceux qui les ont aidés et 2) les individus ne devraient pas nuire à ceux qui les ont aidés (Gouldner 1960, page 171). Dillman (1978) s'appuie sur la théorie des échanges sociaux et en particulier sur la norme sociale relative pour défendre la notion selon laquelle des gestes relativement petits (tel que des lettres personnalisées, des primes



Même si selon un certain nombre d'études récentes, un faible taux de réponse (totale) pourraient ne pas avoir d'effet indésirable grave sur la qualité des données (Curtin, Presser et Singer 2000 ; Keeter, Miller, Kohut, Groves et Presser 2000 ; Visser, Krosnick, Marquette et Curtin 1996), il n'en reste pas moins que la non-réponse totale peut avoir des conséquences négatives sur les estimations statistiques dans certaines circonstances. Par conséquent, trouver des moyens créatifs d'accroître les taux de réponse de manière à ce que tous les types d'unités échantillonnées soient représentés adéquatement dans l'échantillon demeure un objectif clé de la recherche sur les enquêtes. Pour la non-réponse partielle, il est vrai que des techniques applicables après la cueillette des données pour traiter les données manquantes, comme les imputations de type 'hot-deck' ou 'cold-deck', l'imputation par la moyenne, l'imputation multiple combinée ou non à la suppression ont permis d'atténuer les défis que pose ce problème. Cependant, l'idéal serait de réduire le plus possible la non-réponse partielle. En fait, il s'agit là d'un des objectifs principaux de la conception et de la mise en œuvre des enquêtes. Il en est ainsi parce que, dans certains domaines, surtout en microéconomie, la norme consiste à utiliser uniquement les données originales (Cameron et Trivedi 2009).

### 3. Rendre saillantes les caractéristiques clés d'une enquête et participation à l'enquête

La mesure dans laquelle une unité échantillonnée considère comme plus ou moins importantes certaines caractéristiques d'une enquête influe sur la probabilité que cette unité participe à l'enquête (Groves et coll. 2000). Groves et coll. (2000) commentent les techniques d'interview des intervieweurs chevronnés, soulignant que ce que font effectivement ceux-ci quand ils adaptent leurs requêtes ou remarques aux préoccupations des répondants consiste à rendre plus saillantes certaines caractéristiques de la requête, celles qui, selon eux, seront accueillies favorablement par le ménage (page 299). Effortant les travaux de Groves et Couper (1998), Groves et coll. (2000) proposent ce qu'ils appellent la théorie du levier et de la saillance (*leverage-saliency theory*) pour expliquer comment les unités échantillonnées prennent la décision de participer ou de refuser de participer à une enquête. Cette théorie énonce essentiellement que certains attributs (levier) d'une enquête peuvent être considérés négativement ou positivement par le répondant et que la façon dont ces attributs sont rendus saillants durant le processus de demande de participation à l'enquête influe sur la probabilité de participation. Si les attributs perçus positivement par une unité échantillonnée (levier positif) sont rendus saillants durant la demande de participation à l'enquête, il existe plus de chances que la personne

## 2. Non-réponse aux enquêtes et conséquences éventuelles

pratique persistante et très répandue dans le secteur des sondages, nous pensons qu'il appartient aux spécialistes de la recherche par sondage de réévaluer périodiquement la relation entre les primes d'incitation et la participation aux enquêtes, dans des contextes variés. Cette évaluation contribue de l'utilité des primes d'incitation dans les enquêtes est importante, parce que nous ne pouvons presumer que ces primes donneront toujours les résultats attendus.

À la section suivante, nous décrivons brièvement le problème de la non-réponse aux enquêtes, puis nous passons en revue la recherche portant sur la façon dont l'augmentation de la saillance de certaines caractéristiques de l'enquête et l'offre de primes d'incitation prépayées influencent la participation et la non-réponse partielle. Les deux dernières sections traitent de la conception de l'étude et des résultats de cette dernière.

La non-réponse à une enquête décrit la situation où une unité échantillonnée ne participe pas du tout à l'enquête (non-réponse totale) ou ne répond par à une ou à plusieurs questions (non-réponse partielle). La non-réponse aux enquêtes est un problème abordé de longue date dans le domaine de la recherche sur les enquêtes. Singer (2006) observe que, selon une analyse des revues statistiques archivées dans JSTOR, le premier article sur la non-réponse remonte à 1945 et que la référence la plus ancienne dans l'index de la revue *Public Opinion Quarterly* remonte à 1948 (page 637). Cependant, malgré cette sensibilité au problème, tant les projets d'enquête bien établis que les projets naissants connaissent une baisse régulière des taux de réponse. Par exemple, la *Survey of Consumer Attitudes* (SCA) de l'Université du Michigan a vu son taux de réponse passer d'environ 72 % en 1979 à environ 60 % en 1996 et à un creux de 48 % en 2003 (Curtin, Presser et Singer 2005).

La non-réponse aux enquêtes, aussi bien totale que partielle, pose un défi majeur en recherche par sondage, étant donné la possibilité qu'elle introduise des erreurs non dues à l'échantillonnage dans les estimations des paramètres (Brethm 1993 ; Dillman et coll. 2002 ; Groves et Couper 1998). Par exemple, la non-réponse peut donner lieu à des estimateurs ponctuels entachés d'un biais, à un accroissement de la variance des estimateurs ponctuels et à des biais dans les estimateurs de précision (Dillman et coll. 2002 ; Groves et Couper 1998). Bien que la non-réponse totale et la non-réponse partielle soient considérées comme conceptuellement différentes dans les travaux sur les enquêtes, leurs effets sur une estimation statistique sont généralement les mêmes (Groves, Fowler Jr., Couper, Lepkowski, Singer et Tourangeau 2004).



# Examen de la participation aux enquêtes et de la qualité des réponses : l'importance de l'intérêt du sujet et des primes d'incitation

Lazarus Adua et Jeff S. Sharp

## Résumé

Le biais dû à la non-réponse est un problème examiné de longue date dans le domaine de la recherche sur les enquêtes (Brehm 1993 ; Dillman, Eltinge, Groves et Little 2002), et de nombreuses études ont eu pour objectif de préciser les facteurs qui ont une influence sur la non-réponse partielle ainsi que totale. Dans le but de contribuer à la réalisation de l'objectif plus général consistant à réduire au minimum la non-réponse aux enquêtes, nous examinons dans la présente étude plusieurs facteurs pouvant avoir une incidence sur cette non-réponse, en utilisant les données de l'Animal Welfare Survey de 2007 réalisée en Ohio, aux États-Unis. En particulier, nous examinons la mesure dans laquelle l'intérêt du sujet et les primes d'incitation influent sur la participation aux enquêtes et sur la non-réponse partielle, en nous inspirant de la théorie du levier et de la saillance (*leverage-saliency theory*) (Groves, Singer et Comins 2000). Nous constatons que la participation à une enquête est influencée par le contexte du sujet (car celui-ci exerce un effet de levier positif ou négatif sur les unités échantillonnées) et par les primes d'incitation prépayées, ce qui est en harmonie avec la théorie du levier et de la saillance. La constatation que la non-réponse partielle, notre mesure indirecte de la qualité de la réponse, varie en fonction de la proximité par rapport à l'agriculture et l'environnement (lieu de résidence, connaissances sur la production des aliments et options quant à l'importance du bien-être des animaux) confirme aussi nos attentes. Cependant, les données laissent entendre que la non-réponse partielle ne varie pas selon qu'un répondant reçoit ou non une prime d'incitation.

Mots clés : Non-réponse aux enquêtes ; participation aux enquêtes ; levier et saillance ; primes d'incitation prépayées ; non-réponse partielle ; données manquantes.

## 1. Introduction

Le biais dû à la non-réponse est un problème étudié depuis longtemps en recherche sur les enquêtes car il touche tous les travaux effectués dans ce domaine, quel que soit le mode de collecte des données (Nathan 2001). Par conséquent, de nombreux chercheurs ont tenté de déterminer les facteurs qui influent sur la réponse et la non-réponse, partielle ou totale, sous divers modes de collecte (Groves 2006 ; Trussell et Lavrakas 2004 ; Davern, Rockwood, Sherrod et Campbell 2003 ; Teitler, Reichman et Sprachman 2003 ; Singer, Van Hoewyk et Maher 2000 ; Singer, Van Hoewyk, Maher 1998 ; James et Bolstein 1992). Si ces études ont fourni des renseignements éclairants et utiles sur les facteurs qui influencent la participation aux enquêtes, les questions ayant trait à la non-réponse aux enquêtes restent pertinentes dans le domaine de la recherche sur les enquêtes en général et dans nos travaux de recherche fondamentale en particulier. Nous souhaitons poursuivre les réflexions de Groves et coll. (2000) en cherchant à déterminer si des caractéristiques particulières des unités ou sous-populations démographiques échantillonnées en regard du contexte thématique d'une enquête ont une incidence sur les schémas de réponse. Dans le cadre de notre étude permanente des attitudes et comportements du grand public en ce qui a trait au domaine de l'agriculture et de l'environnement, les niveaux constatés de participation à l'enquête et de non-réponse

partielle nous inquiètent de plus en plus. Dans le cas qui nous occupe, l'une des craintes est que la non-réponse totale et partielle pourrait varier selon que les personnes ou les ménages sont plus ou moins proches, physiquement ou socialement, du paysage agricole, qui est le domaine cible de nos sondages d'opinion publique. Afin de contribuer à la réalisation de l'objectif plus général de réduire au minimum la non-réponse totale et partielle, et d'aborder certaines des questions qui nous préoccupent, nous procédons à un nouvel examen de plusieurs facteurs qui ont une incidence sur la participation aux enquêtes et la non-réponse partielle. En particulier, nous examinons les effets du contexte du sujet d'une enquête (c'est-à-dire le domaine cible principal) sur la participation à l'enquête et la non-réponse partielle. Nous nous attendons à ce que la participation à une enquête soit systématiquement influencée par l'intérêt (saillance) que présente le sujet de l'enquête pour chaque unité échantillonnée. Nous nous inspirons à cet égard de la théorie du levier et de la saillance (*leverage-saliency theory*) (Groves et coll. 2000), selon laquelle un ensemble de facteurs liés aux caractéristiques principales d'une enquête ou à des caractéristiques rendues saillantes durant l'administration du questionnaire devrait avoir une incidence sur la participation. Notre étude comprendra aussi le réexamen des effets des primes d'incitation prépayées sur la réponse à l'enquête. Comme l'offre de primes d'incitation aux unités échantillonnées demeure une

- Phippis, P.A., Butani, S.J. et Chun, Y.I. (1995). Research on establishment-survey questionnaire design. *Journal of Business & Economic Statistics*, 13, 337-346.
- Ponikowsk, C.H., et Meilly, S.A. (1989). Controlling response error in an establishment survey. *Proceedings of the Surveys Research Methods Section*. American Statistical Association, 258-263.
- Ramirez, C. (1996). Respondent selection in mail surveys of establishments: Personalization and organizational roles. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 974-979.
- Simon, H. (1957). *Models of man: Social and rational*. New York : John Wiley & Sons, Inc.
- Snijders, G., Onat, E. et Visschers, R. (2007). The annual structural business survey: Developing and testing an electronic form. *Proceedings of the Third International Conference on Establishment Surveys*. *Montreal, Canada*. American Statistical Association, 317-326.
- Sudman, S., Willimiack, D.K., Nichols, E. et Mesenbourg, T.L. (2000). Exploratory research at the U.S. Census Bureau on the survey response process in large companies. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 327-337.
- Tomaskovic-Devey, D., Leiter, J. et Thompson, S. (1994). Organizational survey nonresponse. *Administrative Science Quarterly*, 39, 439-457.
- Tourangeau, R. (1984). Cognitive science and survey methods. *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, (Eds., T.B. Jabine, M.L. Straf, J.M. Tanur et R. Tourangeau), Washington, D.C. : National Academy Press, 73-100.
- Yin, R.K. (2003). *Case Study Research: Design and Methods*. Thousand Oaks : Sage Publications.
- Willimiack, D.K., Nichols, E. et Sudman, S. (2002). Understanding unit and item nonresponse in business surveys. *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dilliman, J.L. Eltinge et R.J.A. Little). New York : John Wiley & Sons, Inc., 213-227.
- Willimiack, D.K., et Nichols, E. (2001). Building an alternative response process model for business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Willimiack, D.K., Lyberg, L.E., Martin, J., Japac, L. et Whitridge, P. (2004). Evolution and adaptation of questionnaire development, evaluation, and testing methods for establishment surveys. *Methods for Testing and Evaluating Survey Questionnaires*, (Eds., S. Presser et coll.). Hoboken : Wiley-Interscience, 385-407.
- Willimiack, D.K. (2007). Considering the establishment survey response process in the context of the administrative sciences. *Proceedings of the Third International Conference on Establishment Surveys*. *Montreal, Canada*. American Statistical Association, 892-903.
- Willimiack, D.K. (2003). Business respondents' perspectives on alternative employment arrangements and implications for employment statistics. *Proceedings of the Section on Government Statistics*. American Statistical Association, 4559-4570.
- Tourangeau, R., Rips, L.J. et Rasinski, K.A. (2000). *The Psychology of Survey Response*. Cambridge, England : Cambridge University Press.

(Sudman et coll. 2000). En outre, le MIRÉE permet la présence d'un effet contagieux de transmission de l'expérience d'une enquête auprès des entreprises à d'autres entreprises auprès des entreprises.

## 5. Conclusion

Les organismes d'enquête doivent généralement réserver une somme considérable de ressources pour le traitement des données, les processus de réponse aux questions de l'enquête dans les entreprises n'étant pas satisfaisants. Le MIRÉE fournit d'autres preuves de la façon dont les processus sont menés et de ce qui les influence. Il donne un aperçu de la perspective des entreprises, qui est utile pour trouver des solutions efficaces en vue d'améliorer les processus et, par conséquent, de réduire ou d'éliminer les erreurs de mesure. Le modèle peut aussi servir de cadre pour la documentation et la systématisation des connaissances existantes et futures des causes des erreurs de mesure dans les enquêtes auprès des entreprises. Il peut être utilisé comme une étape préliminaire des études empiriques des erreurs de mesure et pour une explication uniforme des résultats empiriques. Les recherches à venir devraient inclure l'application de méthodes de recherche qualitative à l'étude des aspects particuliers du processus de réponse, d'autres participants de l'entreprise, mis à part les répondants, et d'autres types d'enquêtes auprès des entreprises. Elles devraient aussi aborder la modélisation quantitative du processus de réponse et vérifier l'efficacité des améliorations suggérées, au moyen d'expériences. Enfin, elles devraient examiner les interactions avec d'autres types d'erreurs non dues à l'échantillonnage.

## Bibliographie

Avison, D., et Elliot, S. (2006). Scoping the discipline of information systems. *Information Systems: The State of the Field*, (Eds., J.L. King et K. Lytinen). Hoboken : John Wiley & Sons, Inc., 3-18.

Bavdaz, M. (2009). Conducting research on the response process in business surveys. *Statistical Journal of the IAO*, 26, 1-14.

## Remerciements

Le présent article est le résultat d'une recherche au niveau du doctorat. L'auteur remercie le Bureau de la statistique de la République de Sloénie pour sa collaboration, ainsi que Lea Bregar (Université de Ljubljana), Lars Lyberg (Bureau de la statistique de la Suède, Université de Stockholm) et Jaak Billiet (Université catholique de Louvain) pour leur aide et leur soutien. Je souhaite aussi remercier le rédacteur associé et les examinateurs anonymes pour leurs commentaires utiles concernant une version antérieure du présent article.

Beatty, P., et Herrmann, D. (2002). To answer or not to answer: Decisions processes related to survey item nonresponse. *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little). New York : John Wiley & Sons, Inc., 71-85.

Biemer, P.P., et Fecso, R.S. (1995). Evaluating and controlling measurement error in business surveys. *Business Survey Methods*, (Eds., B.G. Cox et coll.). New York : Wiley-Interscience, 257-281.

Edwards, W.S., et Cantor, D. (1991). Toward a response model in establishment surveys. *Measurement Errors in Surveys*, (Eds., P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz et S. Sudman). New York : Wiley-Interscience, 211-233.

Giesen, D., et Hak, T. (2005). The response process model in business surveys: Lessons learned by using a multi-method approach. *FCSM Conference Papers*, Federal Committee on Statistical Methodology.

Greenia, N., Lane, J. et Willimaack, D. (2001). Perceptions of confidentiality protection at statistical agencies: Some evidence from data on businesses and households. *Statistical Journal of the United Nations ECE*, 18, 309-314.

Groves, R.M., Fowler, F.J., JR., Couper, M.P., Lepkowski, J.M., Singer, E. et Tourangeau, R. (2004). *Survey Methodology*. Hoboken : Wiley-Interscience.

Hak, T., Willimaack, D.K. et Anderson, A.E. (2003). Response process and burden in establishment surveys. *Proceedings of the Section on Government Statistics*. American Statistical Association, 172-1730.

Hedlin, D., Dale, T., Haraldsen, G. et Jones, J. (2005). *Developing Methods for Assessing Perceived Response Burden*. Eurostat.

Jenkins, C.R., et Dillman, D.A. (1997). Towards a theory of self-administered questionnaire design. *Survey Measurement and Process Quality*, (Eds., L.E. Lyberg et coll.). New York : Wiley-Interscience, 165-196.

Kvale, S. (1996). *Interviews: An Introduction to Qualitative Research Interviewing*. Thousand Oaks : Sage Publications.

Lorenc, B. (2006). Two topics in survey methodology: Modelling the response process in establishment surveys; inference from nonprobability samples using the double sample setup. Thèse de doctorat, Department of Statistics, Stockholm University.

Lorenc, B. (2007). Using the theory of socially distributed cognition to study the establishment survey response process. *Proceedings of the Third International Conference on Establishment Surveys, Montreal, Canada*. American Statistical Association, 881-891.

Morrison, R.L., Seidler, K. et Anderson, A.E. (2002). Using vignettes in cognitive research on establishment surveys. *International Conference on Questionnaire Development, Evaluation and Testing Methods*. American Statistical Association.

Nichols, E.M., Murphy, E.D., Anderson, A.E., Willimaack, D.K. et Sigmam, R.S. (2005). Designing interactive edits for U.S. Electronic Economic Surveys and Censuses: Issues and guidelines. *Research Report Series (Survey Methodology)*, 2005-03). U.S. Census Bureau.

O'Brien, E.M. (2000). Respondent role as a factor in establishment survey response. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 1462-1467.

Statistique Canada, N° 12-001-X au catalogue



(Tourangeau 1984), le MMIREE pousse davantage la notion de connaissance pertinente des processus cognitifs. Williams et Nichols (2001) mettent l'accent sur la connaissance des réponses par les personnes, directement à partir de leur mémoire, et la connaissance des dossiers. Le MMIREE laisse supposer qu'une connaissance approfondie des données des dossiers de l'entreprise et que leur utilisation appropriée pour la réponse à l'enquête nécessitent la connaissance de l'ensemble de la chaîne de production de données, de la connaissance de la réalité de l'entreprise à la connaissance de la constitution des dossiers et à la connaissance des dossiers de l'entreprise.

Pour ce qui est des processus de compréhension, Edwards et Cantor (1991) ont reconnu l'utilisation problématique de jargon, et Sudman et coll. (2000) ont souligné l'écart problématique qui existe entre les concepts économiques requis et les normes comptables. Le MMIREE va encore plus loin et explique que les erreurs peuvent être le résultat d'un problème plus large d'incompréhension des concepts économiques et comptables ou de leur confusion avec d'autres concepts.

Le MMIREE fait état de plusieurs principes qui aident à la compréhension des processus de jugement sous-jacent des enquêtes auprès des entreprises, qui sont conformes aux exemples à l'appui des principes de continuité et de cohérence de Sudman et coll. (2000) et de Williams, Nichols et Sudman (2002), respectivement. Ces principes peuvent aussi rendre compte du caractère satisfaisant des données (Simon 1957) ou de l'inertie. L'utilisation de principes inappropriés, et plus particulièrement le principe de continuité, est principalement renforcée par l'absence de rétroaction d'enquête.

Dans les processus cognitifs de réponse, le MMIREE expose le problème de l'appariement dans les enquêtes auprès des entreprises, ajoutant des éléments à l'erreur d'arrondissement dont il est question dans Sudman et coll. (2000). Il intègre en outre différents aspects de la sensibilité des entreprises dont Edwards et Cantor (1991) ont discuté, dans le cadre de l'étape de communication, de même que Sudman et coll. (2000), dans le cadre de l'étape de diffusion. Le modèle les traite au niveau individuel, là où le contrôle est assuré si les données sont de nature délicate.

#### 4.4 Observations au niveau de l'enquête

Les modèles précédents se sont concentrés sur une seule occurrence du processus de réponse dans une enquête particulière auprès des entreprises, tandis que le MMIREE englobe plusieurs occurrences et plusieurs enquêtes. Parmi les nombreux aspects au niveau de l'enquête, l'étude analytique systématiquement les répercussions de la récurrence et du contact avec le personnel de l'enquête sur le processus de réponse, ce qui constitue une élaboration plus poussée des cas particuliers déjà mentionnés dans des modèles précédents dans le contexte de la recherche, comme la répétition de la consultation (Edwards et Cantor 1991) ou la documentation de la recherche antérieure à l'appui des réponses

approprié, le MMIREE établit un cadre qui peut aussi être utilisé pour la modélisation quantitative et la conception expérimentale.

#### 4.2 Observations au niveau de l'organisation

Les modèles précédents ont abordé les modalités organisationnelles initiales dans le contexte de la sélection des répondants (Biemer et Fecso 1995; Edwards et Cantor 1991) ou dans le cadre d'étapes distinctes de la sélection des répondants et de l'établissement des priorités, ce dernier processus classant la production de données statistiques pour les administrations publiques à un niveau plus bas que la plupart des autres activités de rapports des entreprises (Sudman et coll. 2000; Williams et Nichols 2001). Ils ont aussi déterminé plusieurs facteurs qui influencent la sélection des répondants, et plus particulièrement le rôle fonctionnel, le niveau d'autorité et la position à l'égard du système d'information (Edwards et Cantor 1991), la connaissance du système d'information, des modalités et des définitions (Biemer et Fecso 1995), les responsabilités professionnelles concurrentes et l'accès aux données (Sudman et coll. 2000). Le MMIREE intègre toutes les activités pré-

paratoires dans l'organisation de la réponse à l'enquête et suggère une liste élargie de facteurs d'influence. Dans l'organisation de la réponse à l'enquête, on reconnaît maintenant que la délégation des tâches peut aussi inclure la sélection d'autres participants des entreprises, mis à part les répondants, et que la priorité des tâches concurrentes n'est que l'un des facteurs influençant la planification des tâches.

Dans tous les modèles précédents, on a accordé une attention considérable à la constitution des dossiers. Le MMIREE laisse supposer une systématisation différente et un prolongement des facteurs de la constitution des dossiers, qui ont été regroupées initialement sous les termes gestion, réglementation et normes par Williams et Nichols (2001). Étant donné qu'il est généralement peu probable que les exigences des rapports statistiques constituent un facteur réel de la constitution des dossiers, le MMIREE peut aider l'organisme chargé de l'enquête dans ses tentatives d'influencer la constitution des dossiers et d'obtenir, au bout du compte, les données requises. Le MMIREE, qui tient compte de la capacité technologique et humaine du SIE, définit plusieurs niveaux de disponibilité de réponse selon la mesure dans laquelle la réponse se conforme aux définitions d'enquête requises et propose le résultat de réponse probable. Pour ce qui est de l'autorisation de la réponse dans les entreprises, le MMIREE réitère la possibilité de vérification interne que Sudman et coll. (2000) et Williams et Nichols (2001) proposent pour l'étape de la diffusion. On obtient généralement une autorisation lorsque la réponse à l'enquête fait intervenir des unités juridiquement distinctes et des organisations plus formalisées et centralisées.

#### 4.3 Observations au niveau individuel

Au niveau individuel, c'est-à-dire au niveau de la compréhension, de la recherche, du jugement et de la réponse

L'étude a été axée sur les répétitions de la récurrence du processus de réponse. Lorsque l'enquête est administrée de façon répétée à la même entreprise, l'organisation de la réponse à l'enquête devient moins pertinente ou non pertinente s'il s'agit d'une répétition de l'occurrence précédente. Les processus cognitifs au niveau individuel étaient caractérisés par la routine, lorsqu'ils étaient exécutés par les mêmes participants. De nombreux répondants ont admis qu'ils n'avaient pas lu l'ensemble du questionnaire, sans parler des instructions. Cela s'est aussi produit dans des entreprises qui avaient accepté d'être observées lorsqu'elles remplissaient le questionnaire : une fois que les répondants avaient consulté rapidement le questionnaire pour vérifier s'il comportait des changements, ils se langaient dans le processus de recherche à partir du questionnaire rempli précédemment ou d'autre documentation et des notes d'appui. L'étape de la compréhension a donc été exécutée de façon superficielle et avait davantage trait à la compréhension du questionnaire rempli précédemment qu'à la compréhension des demandes de l'enquête. Les procédures de recherche suivaient la démarche établie précédemment et affichaient des effets de courbe d'apprentissage. Le jugement du répondant se limitait à l'approche initiale et n'était pas susceptible de changer. La récurrence a fréquemment réduit la supervision assurée par le répondant et l'impertinence de l'autorisation, et parfois même menacé son existence.

Étant donné que la tâche de l'enquête était affectée aux mêmes personnes ou services de l'entreprise, bon nombre d'entre eux ont été ou tard eu des contacts avec le personnel de l'enquête, même si l'on avait adopté le mode courant d'auto-administration pour la collecte des données dans les enquêtes auprès des entreprises. Ce contact pouvait se produire tôt dans le processus de réponse et influencer la compréhension et le jugement du répondant. Cela a rarement été le cas dans l'étude ; seulement quelques répondants ont demandé des explications la première fois qu'ils ont participé à l'enquête, et un autre répondant a demandé de l'aide lorsque l'activité de l'entreprise a changé. Les contacts dans le cadre desquels les répondants demandaient un report de l'échéance n'ont pas semblé influencer le processus de réponse subséquent, même si on ne pouvait pas en dire autant pour les répondants qui ont refusé de participer. Tous les autres contacts se sont produits au cours d'un suivi, lorsque le processus de réponse, ou des parties de celui-ci, ont dû être repris, ce qui pouvait donner lieu à une réponse d'enquête rajustée. Même si les répondants ont principalement reconnu la politesse du personnel de l'enquête, leurs appels visaient à signaler que quelque chose allait mal : une échéance manquée, un élément manquant dans le questionnaire, un manque d'uniformité dans les données déclarées. La rareté de ces contacts a fait une impression considérable sur les répondants, parce que ces contacts étaient souvent le seul type de rétroaction assuré par l'organisme statistique. Par contre, les répondants n'ont pas toujours apprécié le manque de rétroaction. Ils s'attendaient à de la rétroaction

#### 4.1 Construction du modèle

La prédominance des communications écrites entre l'organisme chargé de l'enquête et les entreprises a fait en sorte que les participants des entreprises n'étaient plus au centre de la production statistique, ce qui a réduit les possibilités d'intervention dans le processus de réponse aux demandes d'enquête et les causes des erreurs de mesure. En étudiant les mécanismes de réponse et les facteurs d'influence, les modèles de réponse contribuent à faire ressortir ces connaissances et approches en matière de conception, qui font de cette connaissance un avantage. La présente section aborde les contributions du MMIREE aux modèles de réponse précédents s'appliquant aux enquêtes auprès des entreprises.

#### 4. Examen des contributions du modèle

de l'organisme statistique après leur première participation à l'enquête, mais cela ne s'est généralement pas produit. L'absence de réaction les a rendus confiants à l'égard de leur approche, renforçant ainsi le principe de continuité dans leur jugement. Toutefois, de nombreux répondants ont déclaré au moins un élément de donnée qui n'était pas complètement exact (ou pas aussi exact qu'ils l'auraient voulu) et ils ont perçu l'absence de plainte comme une acceptation des données erronées. Certains répondants étaient convaincus que l'organisme statistique connaissait leur activité commerciale, ce qui fait qu'ils ont rarement fourni des descriptions textuelles des mouvements saisonniers. Compte tenu de ces observations, il n'est pas surprenant de constater que plusieurs répondants ont exprimé des doutes concernant l'exactitude des données statistiques, ou remis en question l'exactitude des données fournies par les autres. La rétroaction appropriée n'est pas seulement importante pour cette enquête particulière, mais aussi pour la participation à d'autres enquêtes, parce qu'elle contribue à la perception générale concernant les enquêtes et les statistiques.

Deux approches ont servi à la construction des modèles précédents : l'ajout de certaines étapes organisationnelles ou processus cognitifs de base, à partir du modèle cognitif de réponse aux enquêtes de Tourangeau (Biemer et Fecso 1995 ; Edwards et Cantor 1991 ; Sudman et coll. 2000 ; Williamack et Nichols 2001) ou l'utilisation de l'organisation comme unité d'observation (Lorenz 2006). Le MMIREE relie explicitement les processus et le niveau auquel ils se produisent : les processus cognitifs au niveau individuel et les processus organisationnels au niveau de l'organisation (dans ce cas, l'entreprise). Il prévoit aussi l'observation du processus de réponse sur plusieurs itérations de la même enquête ou sur plusieurs enquêtes comportant des plans différents, ce qui est particulièrement intéressant dans le cas des enquêtes gouvernementales. Grâce à l'analyse des processus de réponse complexes au niveau d'observation



certaines éléments des ventilations des ventes s'ils représentaient moins de 1 % des activités. Les répercussions du principe dépendent de l'utilisation des données recueillies. Il ne devrait pas y avoir de conséquences si l'objectif est d'estimer des variations ou des totaux nationaux. Toutefois, les ventes d'un groupe particulier de produits peuvent être marginales pour une grande entreprise, mais ne pas l'être pour le marché de ce groupe de produits.

Le principe de la perspective de l'entreprise préconise que la perspective de l'entreprise ait priorité sur la demande statistique. Dans l'étude, les données sur les unités organisationnelles existantes ont été considérées comme acceptables, même si elles divergeaient des unités requises ; les données sur les divers types d'emballages (par exemple, un journal accompagné d'un livre) qui étaient pertinentes dans la perspective de l'entreprise n'ont pas été séparées à des fins statistiques.

### 3.3.4 Réponse

La composante de réponse a trait au processus de mise en correspondance d'un jugement et d'une catégorie de réponse, ainsi qu'au contrôle de la réponse (Tourangeau, Rips et Rasinski 2000). Dans les enquêtes auprès des entreprises, la mise en correspondance se traduit habituellement par l'appariement des données disponibles du SIF et des catégories de réponse offertes, ce qui laisse de la place pour une forme particulière d'erreur de mesure : la classification erronée. Par exemple, lorsque les répondants avaient des problèmes à intégrer les données sur les ventes disponibles dans le modèle de classification fourni, ils ont souvent choisi la catégorie la plus proche, la catégorie principale ou la catégorie « autre ».

L'étude a aussi fait ressortir la présence de processus de contrôle montrant les différents aspects de la sensibilité des entreprises. Certains répondants à l'étude ont vérifié si leur sélection du code d'activité définitif correspondait à l'activité enregistrée, ce qui pourrait être le signe qu'ils craignent de ne pas se conformer aux exigences administratives. Le fait de ne pas déclarer des personnes travaillant dans l'entreprise familiale peut être révélateur d'un cas d'évasion fiscale. Même si de nombreux répondants ont convenu que les données déclarées dans le questionnaire étaient considérées comme confidentielles, on a noté peu de preuve de non-divulgation à dessein de données à l'organisme statistique (par exemple, ne pas déclarer de données détaillées sur les nouvelles activités).

### 3.4 Niveau de l'enquête

Le MMIRFEE offre la possibilité de conceptualiser le processus de réponse sur plusieurs itérations d'une enquête ou sur plusieurs années. Il permet donc, au niveau conceptuel, l'observation de la façon dont les éléments du plan d'enquête, qui relève du contrôle de l'organisme chargé de l'enquête, influencent le processus de réponse.

faire le lien entre la connaissance de la réalité de l'entreprise et la connaissance des dossiers de l'entreprise (voir la figure 5). Son importance a été notée, par exemple, au cours de l'observation d'une répondante qui remplissait la questionnaire et qui devait jongler avec l'incohérence des données récupérées sur les ventes. Afin de déterminer les erreurs, elle a dû analyser systématiquement les activités autres que de vente au cours de la période observée et le caractère approprié de leur codage dans les dossiers pour finalement découvrir une transaction qui n'aurait pas dû être incluse dans les chiffres de ventes.

Toutefois, le manque de connaissance n'expliquait pas certains jugements comportant un résultat de réponse défavorable, ce qui fait que l'on a étudié plus étroitement les principes guidant le jugement. Parmi les principes les plus répandus du processus étudié de réponse à l'enquête figurait le principe de continuité, qui préconise l'utilisation de la même stratégie de réponse dans les enquêtes récurrentes, même si cela mène à des erreurs. La continuité a été étudiée au cours d'une même année, mais aussi d'une année à l'autre. Elle semblait s'être améliorée comme en faisait foi l'absence de rétroaction négative de l'organisme statistique et sa satisfaction présument à l'égard des données. L'étude a permis d'identifier plusieurs répondants qui ont utilisé des procédures détaillées de calcul qui étaient assez désuètes. Un répondant a même laissé de côté, par erreur, la section des ventes à commission, mais n'a pas modifié la procédure au cours de l'année, afin d'éviter les ruptures dans les données déclarées.

Deux autres principes ont été déterminés en rapport avec le principe de continuité : le principe d'uniformité et le principe de la non-prise en compte des cas exceptionnels. Le principe d'uniformité signifie l'utilisation de la même stratégie de réponse ou de stratégies de réponse similaires dans le même questionnaire d'enquête. Par exemple, un répondant qui a attribué diverses marchandises à un seul groupe de produits du commerce de gros a fait la même chose pour le commerce de détail. Un répondant qui a estimé le chiffre d'affaires du commerce de gros à partir des données de la TVA a utilisé la même approche pour le chiffre d'affaires du commerce de détail, *etc.* Le principe de la non-prise en compte des cas exceptionnels signifie laisser de côté les nouvelles activités ponctuelles ou temporaires. Par exemple, un répondant à l'étude a déclaré par inadvertance une activité temporaire non comprise dans le questionnaire ; un autre a fait mention de l'exclusion de nouvelles activités des rapports en raison de leur succès incertain. La question est toutefois de déterminer la façon d'établir les limites quant à la nouveauté et au caractère temporaire, ainsi que le moment précis où ces activités deviennent représentatives de l'entreprise.

Le principe de la non-prise en compte des cas exceptionnels est aussi lié au principe de la non-prise en compte des cas marginaux, c'est-à-dire laisser de côté les activités qui sont perçues comme marginales pour l'entreprise. Par exemple, certains répondants à l'étude ont laissé de côté



Cette connaissance est essentielle pour déterminer si les questions de l'enquête s'appliquent à l'entreprise et pour fournir des réponses correctes par la suite. En fait, aucune entreprise visée par l'étude n'a répondu à toutes les questions de l'enquête. Les répondants devaient remplir uniquement les sections qui s'appliquaient au type d'activité commerciale qu'ils avaient. Les questions de l'enquête les ont aussi obligés à sélectionner les groupes de biens et services pertinents, les types d'emploi, les types d'acheteurs du commerce de gros, les types de paiements du commerce de détail, *etc.* La connaissance requise de la réalité de l'entreprise était parfois spécifique : un répondant, par exemple, avait besoin de données concernant les relations entre l'entreprise comme franchisseur et ses franchisés, afin d'éviter le dédoublement en double ou l'exclusion de certains éléments.

Parmi les obstacles majeurs liés à l'utilisation de la connaissance de la réalité de l'entreprise pour bien comprendre les questions de l'enquête figurent la mauvaise compréhension des concepts économiques et comptables ou leur confusion avec d'autres concepts. Par exemple, un répondant avait de la difficulté à faire une distinction entre le concept de commerce, qui comprend le regroupement de biens, et le concept de production, qui comprend une transformation des biens au-delà de leur regroupement ; quelques répondants se sont posés des questions concernant la vente à commission, parce que leur activité était la vente, mais que la comptabilité la traitait comme un service ; de nombreux répondants associaient le commerce de détail à un magasin plutôt qu'à des personnes comme consommateurs finaux, peu importe le type d'acheteur ; un répondant a défini le commerce de gros comme « toute activité commerciale non payée au comptant », plutôt que de le lier à la consommation non finale ; certains répondants ne comprenaient pas que les « organisations non commerciales et autres que de fabrication » étaient des fournisseurs de services ; d'autres ne comprenaient pas la différence entre les marchandises et les matériaux, ces derniers étant un facteur de production (et non pas de commerce) dans la terminologie comptable, mais prenant un autre sens en langage courant, par exemple, les matériaux de construction.

Les répondants à l'étude ont souvent utilisé leurs propres définitions pour interpréter les questions de l'enquête. Il en va de même pour les participants des entreprises qui ont fourni des données sur demande sans consulter le questionnaire et/ou le livret d'instructions. Cela s'est produit par exemple dans quelques entreprises importantes, où les fournisseurs de données dépendaient complètement de leurs propres définitions d'espace de vente lorsqu'ils ont fourni des données sur la répartition des magasins selon la taille de l'espace de vente, des applications plus poussées n'étant disponibles que dans le livret d'instructions.

### 3.3.2 Recherche

Dans les processus de recherche, les données et l'information requises pour répondre à l'enquête sont localisées et

recupérées. Dans les enquêtes auprès des entreprises, les données se trouvent habituellement dans les dossiers de l'entreprise, et non pas dans les mémoires des gens, mais leur interprétation. La recherche dépend donc principalement de la connaissance des dossiers de l'entreprise, qui a trait au contenu et à l'emplacement des dossiers dans l'entreprise et aux possibilités d'accès aux données. Y compris la connaissance des applications et des personnes qui en sont chargées.

Les répondants à l'étude ont principalement affiché une bonne connaissance des dossiers de l'entreprise qu'ils utilisaient. Dans quelques entreprises où les supérieurs ont participé au processus de réponse, ces derniers ne connaissaient pas tous les détails des dossiers et ont demandé à un adjoint d'effectuer la recherche, mais ils avaient une excellente connaissance de la réalité de l'entreprise et savaient comment elle était transposée dans les dossiers. Toutefois, même une connaissance parfaite des dossiers de l'entreprise ne suffisait pas toujours pour récupérer les réponses. Lorsque les dossiers de l'entreprise ne comportaient pas toutes les données nécessaires, la connaissance de la réalité de l'entreprise devenait essentielle pour faire les hypothèses appropriées et produire de bonnes estimations. Cela s'est parfois produit dans de grandes entreprises et des entreprises comptables où les répondants connaissaient très bien les dossiers, y compris le plan comptable et ses codes, mais n'avaient qu'une connaissance vague de l'assortiment de marchandises. Par conséquent, ils ont dû utiliser des estimations pour classer les ventes selon les groupes de produits, leur connaissance des activités de l'entreprise n'étant pas comparable à la connaissance exhaustive et de première main du personnel des ventes. Dans les entreprises plus petites, l'absence des données nécessaires dans les dossiers a parfois signifié qu'il a fallu dépendre complètement de la mémoire des personnes, plutôt que des dossiers ; un répondant, par exemple, est arrivé au chiffre d'emploi pour le commerce de gros en dénombrant les personnes dans les lieux de travail pertinents, à savoir les chauffeurs de camion, les personnes qui travaillaient à l'entrepôt, le personnel des ventes et les commis de bureau.

### 3.3.3 Jugement

Le jugement a trait à la compilation de toutes les données et de l'information récupérées pour formuler une réponse Dans la présente étude, il a souvent trait à une manipulation ou à un traitement des données, comme des sommes, un équilibrage avec les données résiduelles, une nouvelle catégorisation et l'application de proportions. Le jugement est principalement appuyé par la connaissance de la constitution des dossiers. Cette connaissance fournit des renseignements sur la façon dont la réalité de l'entreprise est transposée dans les dossiers et fait en sorte que les données saisies ne sont pas considérées comme des chiffres isolés, des codes ou des mots, mais prennent une certaine signification représentant les processus et les objets mesurés. Il sert par conséquent à

modèle de réponse de Tourangeau (1984). Dans les enquêtes auprès des entreprises, ces processus ne sont pas définis aussi facilement que dans les enquêtes auprès des particuliers, parce qu'il arrive que l'organisation initiale de la réponse fasse intervenir uniquement un examen bref et superficiel de la tâche de réponse, ce qui n'a à peu près pas de répercussions sur le processus de réponse ultérieur, ou encore une réflexion approfondie au sujet des questions. L'étude a été axée principalement sur les processus cognitifs des répondants parce que c'est à eux qu'il revient de répondre aux questions d'enquête. Néanmoins, des observations des autres participants des entreprises sont fournies lorsqu'elles sont disponibles.

### 3.3.1 Compréhension

Dans les processus de compréhension, les répondants interprètent les demandes de données de l'enquête, qui prennent habituellement la forme d'étiquettes plutôt que de questions. Selon le MMIRÉE, dans le cas des processus de compréhension, la connaissance de la réalité de l'entreprise est particulièrement importante. La réalité de l'entreprise a trait aux activités que mène l'entreprise pour subsister et à la répartition du travail entre les emplacements et les personnes. La connaissance de la réalité de l'entreprise suppose donc une connaissance de tous les aspects de l'entreprise : qui fait quoi, quelles sont les activités de l'entreprise et comment sont-elles menées, comment les décisions sont-elles prises, pourquoi la situation de l'entreprise est-elle ce qu'elle est, comment a-t-elle évolué au fil du temps, etc. Étant donné que les entreprises plus importantes ont tendance à être complexes et à comporter des divisions techniques et sociales du travail, des filiales, une hiérarchie organisationnelle et une structure de prise de décisions (Tomaskovic-Devey, Leiter et Thompson 1994), on peut s'attendre à ce que la fragmentation de la connaissance de la réalité de l'entreprise augmente avec la taille de l'entreprise.

comprises dans la présente étude ont trouvé cette étape de l'organisation inutile et l'ont même sautée. Dans plus de la moitié des entreprises, les répondants ont signé les questionnaires eux-mêmes, parce « qu'ils avaient le mandat de signer ce genre de choses » et parce que « le directeur est très rarement sur place » ou « ne s'occupe pas de ces choses ». Toutefois, même dans ces cas, certains répondants ont mentionné que le directeur avait été informé de la procédure. Dans plusieurs entreprises, le supérieur a signé le questionnaire pour l'officialiser et aucune procédure de vérification n'était en place, parce que « le directeur faisait confiance » ou « n'avait pas les données nécessaires », ou parce que « nous travaillons de cette façon ». Un supérieur était généralement présent dans les entreprises les plus importantes et était à la source d'une autorisation officielle ou d'un avis informel. Les cas de vérification interne étaient rares, ce qui pourrait être une conséquence des consultations avec le supérieur ayant précédé. Les entreprises comptables fournissaient habituellement le questionnaire rempli à l'entreprise pour qu'il soit signé, même si certains responsables d'entreprise ont parfois signé le questionnaire non rempli à l'avance.

### 3.3 Niveau individuel

Compte tenu du niveau de disponibilité des réponses dans le SIF, c'est le résultat des processus cognitifs et des mesures matérielles qui les accompagnent (particulièrement l'interaction avec les ordinateurs) au niveau individuel qui détermine le résultat final de la réponse. Selon le MMIRÉE, trois types de connaissances intrinsèquement liées sont pertinents pour ces processus : connaissance de la réalité de l'entreprise, connaissance de la constitution des dossiers et connaissance des dossiers de l'entreprise (voir la figure 5). Même s'il peut être difficile de séparer les trois types de connaissances dans les faits, l'étude semble supposer que chaque type a une influence particulière sur un type de processus cognitif.

La répartition des processus cognitifs entre la compréhension, la recherche, le jugement et la réponse est tirée du

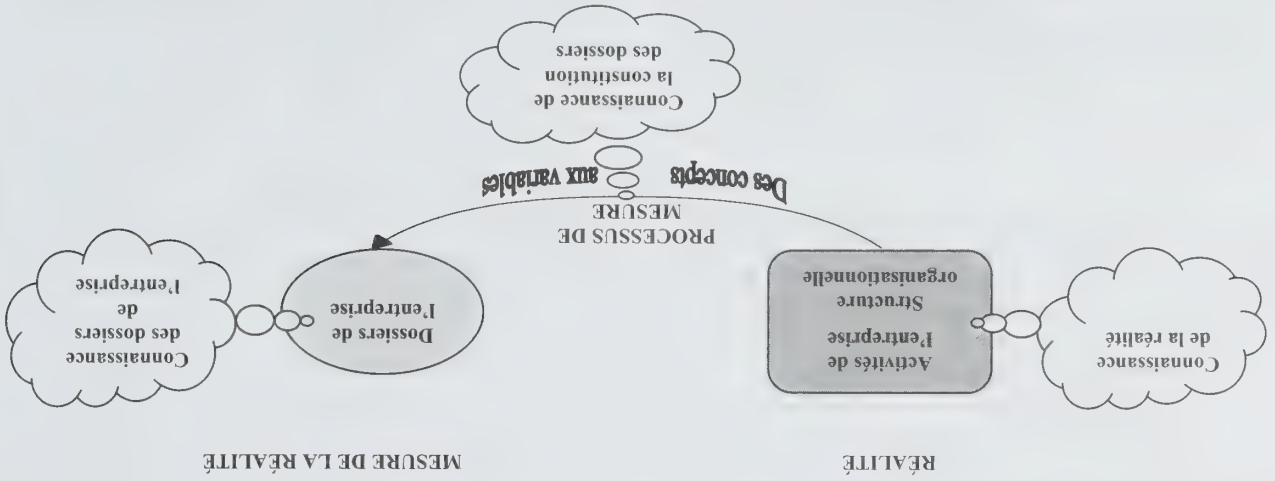


Figure 5 Connaissances pertinentes au processus de réponse aux enquêtes auprès des entreprises



leur conformité aux définitions de l'enquête. La disponibilité des données semble se situer à l'intersection de la capacité technologique et humaine dans l'entreprise; des connaissances sont requises pour extraire des données du SIF, à condition que de telles données existent. Plusieurs niveaux de disponibilité de réponse dans le SIF s'appliquent aux questions d'enquête (voir la figure 4); leur désignation a été inspirée par la détermination des états cognitifs dans Beatry et Hermann (2002) et est en principe conforme à ce qui est proposé par Lorenc (2007) :

- Les données sont accessibles – la réponse requise est facilement disponible. Dans cette étude, un exemple type est le total des revenus des ventes, qui est facilement disponible auprès des responsables de la comptabilité, ou sur le nombre d'employés, qui est facilement disponible auprès du service du personnel.
- Les données peuvent être produites – la réponse requise n'est pas facilement disponible pour quiconque; les données disponibles représentent la base de la production de la réponse requise, grâce à la manipulation. Dans l'étude, par exemple, les revenus de ventes d'une activité commerciale particulière n'étaient pas toujours facilement disponibles, mais il était possible de calculer le chiffre exact en consultant deux dossiers distincts (c'est-à-dire le grand livre général et les dossiers commerciaux).
- Les données peuvent être estimées – la réponse requise n'est pas facilement disponible pour quiconque; les données disponibles représentent la base de la production de la réponse requise, grâce à la manipulation. Dans l'étude, une ventilation des ventes selon les groupes de produit (par exemple, aliments, boissons, vêtements, chaussures) a souvent été estimée en répartissant à nouveau entre des catégories les groupes disponibles; toutefois, ces catégories étaient parfois trop agrégées ou trop diversifiées pour permettre un appariement exact (par exemple, les articles de Noël, les cadeaux de Pâques, les produits discontinus).
- Les données sont inconcevables – aucune donnée disponible ne mène à la réponse requise ou à une approximation de celle-ci; certaines bases pour la production ou l'estimation de la réponse requise existent, mais elles nécessitent un effort imaginable pour la produire. Par exemple, une compagnie devrait classer plus de 10 000 factures chaque mois pour produire une ventilation exacte des ventes selon le type d'acheteur.
- Les données sont inexistantes – il n'existe pas de base pour l'estimation de la réponse requise. Dans l'étude, un magasin libre-service ne peut pas faire de distinction entre les différents types d'acheteurs,

parce qu'il délivre le même type de facture non identifiée à tous les clients, compagnies et particuliers. Comme la disponibilité des données varie selon les personnes appartenant à l'entreprise, il peut être utile de déterminer la disponibilité des réponses au niveau individuel. Dans ce cas, une distinction doit être faite entre une réponse que quelqu'un peut obtenir directement et une réponse qui est accessible uniquement par l'entremise d'une autre personne.

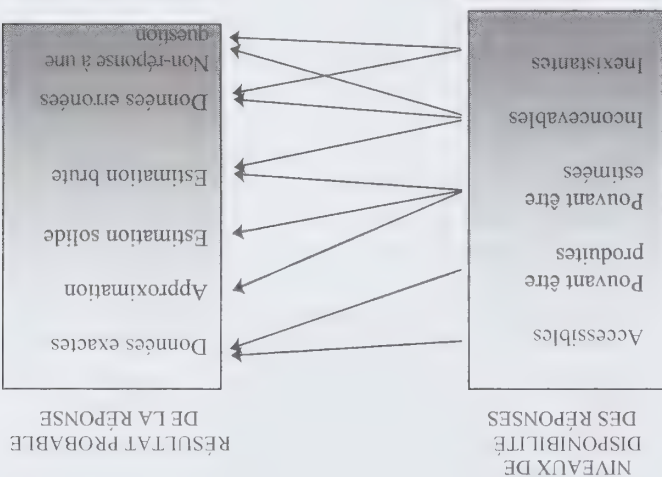


Figure 4 Niveaux de disponibilité des réponses et résultat probable de la réponse

Le résultat de la réponse finale est conditionnel au niveau de disponibilité des réponses et peut aller de données exactes à la non-réponse à une question particulière (voir la figure 4). Une erreur de mesure se produit lorsque le résultat de la réponse s'écarte des données exactes. Lorsque les données sont accessibles ou qu'elles peuvent être produites, la réponse aura probablement pour résultat des données exactes, même si la possibilité de commettre une erreur augmente si les données sont accessibles par l'entremise d'autres personnes ou doivent être manipulées. Lorsque les données peuvent être estimées, le résultat de la réponse peut constituer une approximation comportant une erreur de mesure négligeable ou une estimation comportant une erreur de mesure mineure ou substantielle. Les données inconcevables peuvent, au mieux, mener à une estimation brute. Lorsque les répondants ne disposent pas de base appropriée pour fournir une réponse, ils peuvent faire des suppositions tous azimuts, qui donnent lieu à des erreurs, ou ils peuvent sauter la question, ce qui entraîne la non-réponse à une question.

### 3.2.3 Autorisation de la réponse dans l'entreprise

L'autorisation constitue la dernière occasion de prendre des mesures correctives avant que la réponse de l'entreprise ne soit envoyée à l'organisme chargé de l'enquête et que la documentation ne soit archivée. La plupart des entreprises



activités économiques similaires. Ils fournissent un cadre dans lequel les compagnies élaborent leurs propres solutions internes, à moins que le respect des exigences obligatoires ne réussisse à combler entièrement les besoins de données nécessaires au fonctionnement de l'entreprise, comme c'était le cas dans les petites compagnies locales. Parmi les facteurs internes de la constitution des dossiers figurent les caractéristiques de l'activité de l'entreprise, par exemple, l'importance de ses activités, leur type et leur diversité ; l'intégration dans l'environnement commercial ; et la disposition à constituer des dossiers.

L'importance des activités de l'entreprise joue un rôle crucial dans la constitution des dossiers, parce qu'elle fournit un aperçu différent. Dans l'étude, la plupart des grandes entreprises disposaient d'une abondance de données. Les dossiers des entreprises fournissent des renseignements qui ne peuvent pas être obtenus par la participation ou l'observation seulement. Ceci étant dit, l'importance des activités de l'entreprise est relative, particulièrement si la taille est observée uniquement à l'intérieur des limites juridiques ou des frontières nationales. Par conséquent, il vaut mieux parler de l'intégration dans les réseaux de divers types. Dans l'étude, par exemple, quelques entreprises plus petites étaient sous contrôle étranger, ce qui exigeait la production de rapports exhaustifs, malgré la distance, et la gestion de l'entreprise à distance, et une autre petite entreprise devait utiliser le logiciel perfectionné d'un partenaire commercial, parce que celui-ci était son principal fournisseur. L'étude a aussi fait ressortir comment les différents types d'activités ont influencé le type de dossiers disponibles ; par exemple, les entreprises de commerce de gros qui indiquent habituellement le nom de leurs destinataires sur leurs factures avaient davantage d'information sur leurs acheteurs que les entreprises du commerce de détail, qui fournissent habituellement des reçus sans indiquer de nom. La grande diversité des activités commerciales représente aussi un défi de taille pour la constitution des dossiers dans la plupart des entreprises ; en général, les entreprises plus petites avaient renoncé à utiliser des dossiers détaillés et étaient forcées de produire des estimations à la place. Enfin, la disposition à trait aux attitudes des personnes de l'entreprise à l'égard des divers aspects de la constitution des dossiers, par exemple, le désir de produire des données, la technologie de l'information et les changements. Certaines entreprises dépendaient dans une large mesure d'un processus de décision fondé sur des faits probants et avaient une haute opinion des données ; d'autres ont manifesté de l'enthousiasme à l'égard des possibilités de la technologie de l'information, mais quelques autres ne voyaient aucune utilité aux données.

Les facteurs de la constitution des dossiers influencent la disponibilité des données dans les dossiers de l'entreprise et

exigences obligatoires figurent les contributions, les garanties, l'assurance, les questions environnementales, etc. Les participants ont habituellement noté le caractère obligatoire des enquêtes gouvernementales auprès des entreprises, même si l'absence de sanctions en cas de non-réponse ou de réponse tardive en a incité certains à le remettre en question ; par ailleurs, la modification de la constitution des dossiers à des fins statistiques était impensable pour la plupart des participants à l'étude. Les normes représentent un aspect moins rigide des facteurs externes : elles ne sont pas obligatoires, mais on s'attend à ce qu'elles soient appliquées dans la plupart des cas. Deux exemples tirés de l'étude comprennent l'utilisation d'une classification fondée sur le code à barres du numéro d'article européen et les recommandations des responsables comptables. Selon l'étude, les normes n'étaient pas appliquées pour des raisons particulières ; par exemple, les systèmes d'information des détaillants les plus petits n'utilisaient pas de codes à barres. Les pratiques repères sont le groupe de facteurs externes qui le moins d'influence. Elles ont trait à des exemples de bonnes pratiques qui ont acquis une certaine reconnaissance et autorité, de par leur réputation (et non pas selon la loi ou un pouvoir institutionnel). Par exemple, certains répondants à l'étude ont mentionné les logiciels desuets par rapport aux normes courantes, tandis que d'autres ont souligné la puissance de leur logiciel et son influence positive sur la fourniture des données.

FACTEURS INTERNES FACTEURS EXTERNES

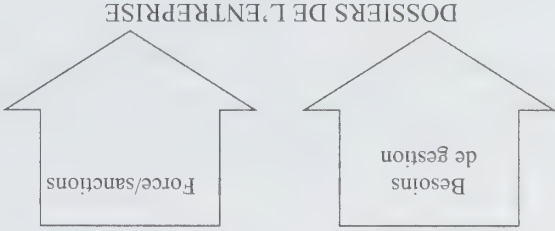
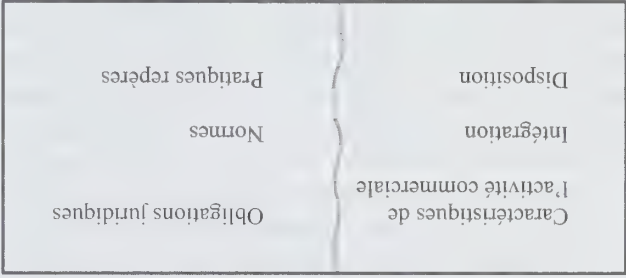


Figure 3 Facteurs liés à la constitution des dossiers

Des facteurs externes sont à la base de l'homogénéité des données et de la comparabilité des dossiers d'une entreprise à l'autre, à tout le moins dans le cas de celles qui ont des

La concurrence entre les tâches est liée à l'affectation des personnes et à l'ordre des tâches. Elle a généralement une influence sur le choix des participants des entreprises au niveau organisationnel, lorsque divers participants possibles sont comparés, ainsi que sur la planification de la tâche de réponse à l'enquête au niveau individuel, lorsque les priorités des diverses tâches d'un participant sont prises en compte. Dans plusieurs petites entreprises principalement, les répondants à l'étude ont convenu qu'ils accordaient un faible niveau de priorité à la tâche de réponse aux enquêtes lorsqu'ils planifiaient leur travail : « la TVA (taxe sur la valeur ajoutée), le recouvrement des dettes, la tenue de livres... ont tous la priorité sur les statistiques ». Une autre répondante a indiqué qu'elle « ne penserait jamais répondre à l'enquête le jour où tous les enregistrements comptables sont effectués », mais qu'elle vérifie plutôt « le bilan... les passifs, l'état des paiements, le montant de la dette, la situation financière ». Un autre a expliqué le déroulement du travail de la façon suivante : « rapports internes en premier lieu, affaires courantes par la suite, rapports statistiques après ». Dans quelques grandes entreprises, toutefois, les répondants ont indiqué qu'ils remplissaient les questionnaires d'enquête dès que les données étaient disponibles ou finales.

De même, les attitudes à l'égard de la tâche de réponse aux enquêtes peuvent être examinées au niveau organisationnel, par l'entremise des politiques officielles sur les enquêtes et des réactions informelles des autorités, ainsi que des perceptions individuelles. Les entreprises visées par la présente étude n'avaient pas de politique officielle sur les enquêtes, même si le discours tenu par les autorités dans certaines entreprises faisait ressortir leurs attitudes négatives : « C'est seulement des statistiques ; préparez quelque chose ». Les attitudes de l'organisation peuvent avoir des répercussions sur l'organisation de la réponse aux enquêtes, qui prennent la forme de conséquences possibles pour l'entreprise, et plus particulièrement, de coûts d'opportunité, de sanctions et de mauvaise image dans le public. La plupart des participants ont exprimé une attitude négative à l'égard des enquêtes, les décrivant comme « un mal nécessaire » et comme « redondantes » ou « une source de travail additionnelle ». Les attitudes individuelles à l'égard des enquêtes peuvent contribuer à la planification précocce, en temps opportun ou tardive de la tâche ; elles peuvent aussi influencer l'inclusion d'une personne dans la tâche de réponse ou son exclusion.

La constitution de dossiers et la fourniture de données sont des éléments clés de la planification des tâches de réponse. L'échéancier de la constitution des dossiers sert à déterminer quand les dossiers comportant les données requises concernant l'entreprise ont été créés dans une forme acceptable ou souhaitable, particulièrement lorsque les

### 3.2.2 Récupération de données dans le système d'information de l'entreprise

La capacité du système d'information de l'entreprise (SIE) est le facteur clé qui influence le processus de réponse et son résultat dans les enquêtes auprès des entreprises. Le SIE ne comprend pas uniquement l'élément technologique ; il englobe aussi les personnes (Avison et Elliot 2006). La capacité humaine du SIE pertinente pour la réponse aux enquêtes auprès des entreprises se reflète principalement dans les processus cognitifs au niveau individuel (voir la section 3.3), tandis que sa capacité technologique est déterminée grâce aux dossiers de l'entreprise au niveau organisationnel. L'étude a démontré que la constitution des dossiers de l'entreprise dépend de facteurs internes et externes, ce qui fait que la limite entre les deux groupes est floue (voir la figure 3).

Des facteurs externes – obligations juridiques, normes et pratiques répérées – sont imposés aux compagnies par l'environnement et dictent de force ou par la menace de sanctions le contenu des dossiers des entreprises. Les dispositions législatives, les règlements et d'autres formes de pouvoirs prévus dans la loi définissent les obligations juridiques. À cet égard, les répondants de l'étude ont mentionné principalement le respect obligatoire des normes comptables et des exigences fiscales. Ces derniers éléments se rapportent à l'ensemble de l'entreprise (par exemple, les rapports sur la TVA) ou à des éléments précis (par exemple, les taxes d'accise sur les produits du tabac). Parmi les autres

Une fois la tâche de réponse à l'enquête organisée, elle peut être exécutée, même s'il est parfois nécessaire de préciser d'avantage la sélection des participants des entreprises ou la planification, en vue de fournir tous les éléments requis, et de prévoir les absences du travail et d'autres circonstances.

consignation des factures qui entrent. rapport aux chiffres de ventes, en raison de retard dans la plus de temps pour obtenir la valeur correcte des stocks, par certains répondants ont expliqué qu'ils avaient besoin de dans les faits la planification proprement dite. Par exemple, dossiers et la dernière fourniture de données déterminent données requises, ce qui fait que la dernière constitution de la fourniture des données peut varier selon le type de Toutefois, l'échéance liée à la constitution des dossiers et à la fourniture des données peut varier selon le type de données requises. Cela s'applique particulièrement aux entreprises comptables dans la présente étude. Les cas où le participant dépend d'autres personnes pour fournir les données requises. Cela s'applique particulièrement aux entreprises comptables dans la présente étude. Toutefois, l'échéance liée à la constitution des dossiers et à la fourniture des données peut varier selon le type de données requises, ce qui fait que la dernière constitution de dossiers et la dernière fourniture de données déterminent dans les faits la planification proprement dite. Par exemple, certains répondants ont expliqué qu'ils avaient besoin de plus de temps pour obtenir la valeur correcte des stocks, par rapport aux chiffres de ventes, en raison de retard dans la consignation des factures qui entrent.



3.2 Niveau organisationnel

3.2.1 Organisation de la réponse à l'enquête

La participation à une enquête nécessite habituellement certaines activités préparatoires, en raison de la répartition du travail et de la spécialisation dans les organisations. Elle requiert la désignation de la personne qui s'occupera de répondre à l'enquête et du moment où cela se fera ; ces deux éléments fournissent des indices concernant la façon dont la tâche sera menée. L'étude a fourni des preuves que les deux étapes peuvent être liées de façon intrinsèque. En fait, la sélection des personnes devant répondre à l'enquête peut en soi indiquer la priorité accordée à la tâche dans l'organisation. Par exemple, dans certaines entreprises comptables et entreprises importantes, les chefs s'acquittent de la tâche eux-mêmes, même s'ils auraient pu la déléguer, ce qui peut indiquer que la tâche revêt une certaine importance, tandis que le fait qu'elle soit confiée à des répondants novices peut indiquer son faible niveau de priorité. En comparaison, les priorités au niveau individuel n'étaient pas toujours conformes aux priorités au niveau organisationnel. Par exemple, même si les déclarations de revenu ont acquis une priorité plus grande que les données statistiques au niveau organisationnel, cet élément n'est pas pertinent pour un répondant d'enquête qui ne produit pas de données fiscales. J'examine par conséquent la sélection des participants des entreprises et la planification de la tâche de réponse en même temps que l'organisation de la réponse à l'enquête. Cela donne lieu à une liste élargie de facteurs pouvant influencer l'organisation de la tâche de réponse à l'enquête (voir la figure 2).

La tradition, les pratiques courantes, les procédures établies et l'emplacement de l'information influencent la sélection des participants des entreprises, qui est une question organisationnelle, tandis que d'autres facteurs se situent à la fois aux niveaux organisationnel et individuel. La tradition veut que l'on ait recours aux participants antérieurs des enquêtes récurrentes, c'est-à-dire la participation répétée des mêmes personnes au processus de réponse à la même enquête (longitudinale). Certains répondants de l'étude ont prétendu qu'ils y répondaient depuis des années. Certains y

répondent depuis qu'ils ont commencé leur emploi ou depuis le départ à la retraite, le départ pour un congé de maladie prolongé ou la cessation d'emploi d'un collègue. De nombreux processus organisationnels reposent sur des pratiques courantes et des procédures établies, qui entraînent la sélection des participants habituels. Cela signifie que même lorsqu'une nouvelle demande d'enquête est soumise à l'entreprise, ses responsables procéderont probablement de la même façon que pour les demandes d'enquête précédentes, en raison de la répartition relativement stable du travail. En fait, certains des répondants de la présente étude ont expliqué que le questionnaire d'enquête est souvent achevé au même service ou à la même personne, qui y donne habituellement suite, même s'il n'existe pas de politique officielle sur les enquêtes. Comme l'a expliqué un répondant : « Ils préfèrent me les confier – c'est la seule politique en place ». Certains répondants savaient quels types d'enquêtes leur étaient confiées et disaient, par exemple : « Je m'occupe de toutes les statistiques, sauf la paie » ou « Je m'occupe de toutes les statistiques, y compris celles pour la Banque de Slovaquie, mais pas pour Intrastat ». Même dans les grandes entreprises, il arrivait souvent que la même personne réponde à plusieurs questionnaires d'enquête différents ; une personne répondait à tous les questionnaires d'enquête servant à recueillir des données financières, que ce soit pour la Banque de Slovaquie, le Bureau de la statistique ou l'Agence des dossiers juridiques publics, d'autres ont fourni une liste d'enquêtes particulières auxquelles elles devaient répondre, par exemple, des enquêtes sur les investissements, les actifs fixes, la valeur ajoutée, etc. L'emplacement de l'information représente un facteur essentiel qui influence la sélection des participants des entreprises, dans le contexte des erreurs de mesure. Elle présuppose une connaissance suffisante pour fournir une réponse précise, y compris l'accès approprié aux dossiers, au besoin. Dans la présente étude, de nombreux répondants ont indiqué qu'ils avaient été choisis parce qu'ils avaient accès aux données, par exemple : « J'ai les données et je sais comment les récupérer ».

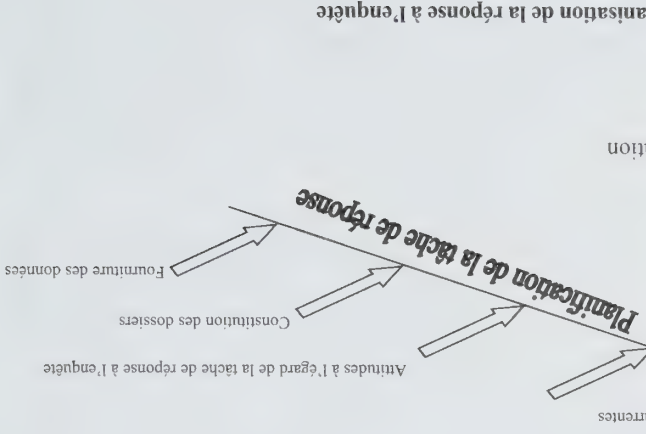


Figure 2 Facteurs qui influencent l'organisation de la réponse à l'enquête



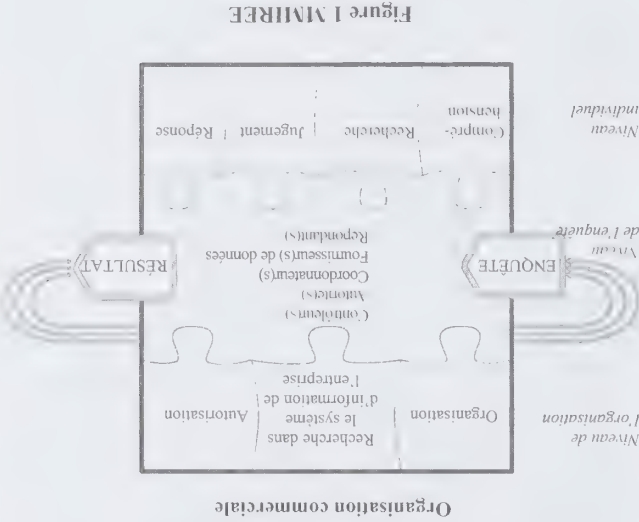
imminentes ayant été annoncées à l'avance. Dans les questions d'interview, on a demandé aux répondants d'indiquer comment ils avaient rempli le questionnaire la dernière fois (par exemple, à la fermeture des livres pour le mois, temps consacré à la réponse, signataire du formulaire et rapidité), et les répondants ont généralement étayé leurs rapports au moyen de données tirées de documents sur papier et électroniques, qu'ils ont utilisés pour remplir le questionnaire. Tout cela a contribué à faire une distinction entre leur dernière participation et leur participation habituelle.

L'interview comme principale méthode de recherche a parfois été combinée à l'observation. Les interviews ont été enregistrées et transcrites. Un plus grand nombre de modèles répétés ont émergé au fur et à mesure que les travaux sur le terrain ont progressé, même si on a noté une baisse des résultats de chaque visite consécutive sur place, vers la fin des travaux sur le terrain. Les résultats des visites sur place ont été comparés aux observations du personnel chargé de l'enquête et des experts spécialisés, aux données quantitatives (lorsqu'elles étaient disponibles) et aux recherches publiées précédemment. D'autres explications ont été envisagées. Enfin, la sélection d'une enquête typique auprès des entreprises a rendu plus plausible la généralisation à d'autres enquêtes auprès des entreprises. Comme le suggère Yin (2003), toutes les étapes de la recherche ont été documentées avec soin, afin d'établir une chaîne de données probantes et d'assurer la fiabilité des résultats.

### 3. MMIREE

### 3.1 Présentation du modèle

L'un des principaux résultats de l'étude est le modèle multidimensionnel intégral de réponse aux enquêtes auprès des entreprises (MMIREE), qui intègre les résultats des recherches antérieures et les nouvelles conclusions de mon étude empirique. Le MMIREE fait une distinction explicite entre les processus qui se produisent au niveau individuel et les autres qui se produisent au niveau de l'organisation, c'est-à-dire au niveau de l'entreprise dans ce cas (voir la figure 1). Les processus cognitifs de compréhension, de recherche, de jugement et de réponse qui ont lieu au niveau individuel sont tirés du modèle de réponse de Tourangeau (1984). Ils rendent compte du processus mental des personnes qui répondent aux enquêtes, en rapport avec la réponse réelle aux questions d'une enquête particulière, comparativement aux processus qui se rapportent à l'organisation, au soutien de l'information et à l'autorisation de la réponse, qui se produisent au niveau de l'entreprise. Contrairement à la situation typique des enquêtes auprès des particuliers, des parties du processus, comme l'obtention de données d'un autre participant ou la récupération de données dans les dossiers de l'entreprise, sont perceptibles



(rôles dans la figure 1). La tâche de réponse aux enquêtes peut faire appel à plusieurs représentants de l'entreprise, qui peuvent intervenir ou cesser d'intervenir dans le processus de réponse à divers moments ; toutefois, pour plus de clarté et de simplicité, ils sont tous regroupés. Les participants des entreprises prennent part aux processus organisationnels tout en maintenant leurs propres processus cognitifs : ainsi, ils constituent un lien unificateur entre les processus aux niveaux individuel et organisationnel. Ils peuvent adopter un ou plusieurs des rôles, qui ont une influence différente sur le processus de réponse, à savoir comme contrôleur (par exemple, réceptionniste, service commun), autorité, coordinateur de la réponse, fournisseur de données ou répondant. Même si la figure 1 présente les participants d'une seule organisation commerciale, l'accomplissement de la tâche peut nécessiter la participation de sous-traitants ou la communication avec le personnel d'enquête.

enquêtes auprès des entreprises : le modèle de réponse hybride pour les enquêtes auprès des établissements de Sudman, Williamack, Nichols et Miesenbourg (2000) et le modèle complet de Williamack et Nichols (2001). Plus récemment, Lorenc (2006) a proposé l'examen de l'en-semble du processus de réponse, selon le concept de la cognition sociale distribuée, l'établissement servant d'unité d'observation.

Ces modèles font ressortir de nombreux aspects essentiels du processus de réponse aux enquêtes auprès des entreprises et présentent certains concepts à cet égard, mais ils n'abordent que partiellement de nombreuses questions. Cela a un effet incitatif pour la tenue d'une étude exhaustive du processus de réponse à une enquête donnée auprès des entreprises et rend possible l'élaboration plus approfondie du modèle de réponse aux enquêtes auprès des entreprises. Le présent article présente le modèle multidimensionnel intégral de réponse aux enquêtes auprès des entreprises (MMIRÉE) et aborde ses contributions.

## 2. Étude empirique

L'objectif de l'étude empirique était d'élaborer un cadre conceptuel du processus de réponse – un modèle de réponse – grâce à l'examen complet du processus proprement dit de réponse à une enquête typique auprès des entreprises, dans un contexte opérationnel réel. L'interview de recherche qualitative a constitué la principale méthode de recherche. La méthode a été mise en œuvre au moyen de diverses techniques (principalement l'approfondissement rétrospectif et les interviews ethnographiques, mais aussi les réflexions à haute voix) et de deux modes (sur place et par téléphone), et auprès des différents répondants (représentants des entreprises participantes, experts de l'administration des questionnaires de l'organisme statistique et experts spécialisés). Dans certains cas, l'observation sur place et des analyses des microdonnées ont complété ces techniques. Compte tenu de toutes les variables, une gamme d'approches ont dû être élaborées (pour plus de détails, voir Bavdaz 2009). Des visites sur place ont été organisées en fonction de deux échéances consécutives pour la réponse au questionnaire en 2005. On a tenté de communiquer avec toutes les personnes clés participant au processus de réponse.

L'enquête choisie – l'Enquête trimestrielle sur le commerce – est une enquête auprès des entreprises menée par le Bureau de la statistique de la République de Slovénie, auprès d'un échantillon d'environ 1 600 entités légales ayant des activités commerciales. L'enquête comportait les caractéristiques classiques des enquêtes auprès des entreprises, c'est-à-dire qu'il s'agissait d'une enquête gouvernementale obligatoire recouvrant par la poste, reposant sur un questionnaire sur papier de huit pages et des livrets d'instructions

et de classification. Le questionnaire comprenait un texte de présentation et quatre sections, une se rapportant à l'en-semble de l'entreprise, et les trois autres, à un type d'activité commerciale (vente à commission, vente en gros et vente au détail). Toutes les sections comportaient des questions sur les ventes et l'emploi. Des questions portaient en outre sur les ventilations des ventes, les stocks, les codes d'activité et le nombre de magasins, ainsi que leur taille. Les unités non répondantes ont reçu jusqu'à trois rappels et, en dernier ressort, un appel téléphonique. Les taux de réponse finaux étaient généralement élevés, c'est-à-dire supérieurs à 90 %. Les principaux écarts et incohérences découverts pendant les procédures de contrôle ont aussi nécessité des appels téléphoniques aux entreprises.

L'échantillon final de cette étude comprenait 28 entreprises qui ont dû répondre à l'Enquête trimestrielle sur le commerce. Les études antérieures ayant donné lieu à des modèles de processus de réponse s'appliquant aux enquêtes auprès des entreprises étaient aussi fondées sur de petits échantillons : 24 établissements (Edwards et Cantor 1991), 30 compagnies comportant plusieurs unités (Sudman et coll. 2000 ; Williamack et Nichols 2001), et sept écoles (Lorenc 2006). Cela va dans le sens des études par interviews exploratoires qui ont tendance à avoir des échantillons de petite taille d'« environ  $15 \pm 10$  » (voir Kvale 1996, page 102). La sélection des entreprises visait à tenir compte de l'hétérogénéité des processus de réponse. Étant donné que la taille de l'entreprise peut être définie comme l'une des caractéristiques les plus importantes de l'entreprise qui a une influence sur les caractéristiques du processus de réponse ou est liée à ces caractéristiques (par exemple, O'Brien 2000), des entreprises ont été sélectionnées dans toutes les catégories de taille.

Plusieurs mesures ont rehaussé la validité du plan de recherche. Les entreprises ont été sélectionnées à partir de différentes catégories de taille, y compris certaines des plus importantes dans le secteur du commerce, mais aussi certaines dont la principale activité n'était pas commerciale. Quelques entreprises ont refusé de collaborer, en raison principalement d'une surcharge de travail. Néanmoins, il faut faire preuve de prudence lorsqu'on applique les constatations aux entreprises non commerciales et aux entreprises surchargées. L'étude a inclus des personnes jouant différents rôles dans le processus de réponse. Des efforts substantiels ont été déployés pour susciter la participation et organiser des visites au cours de la période pendant laquelle les répondants remplissaient le questionnaire, ou juste après, afin de réduire la possibilité qu'ils oublient des renseignements. Les petits décalages dans le temps qui se sont produits dans certains cas n'ont pas semblé avoir affecté la capacité de se rappeler un processus fréquemment répété et bien documenté, les visites sur place



# Modèle multidimensionnel intégral de réponse aux enquêtes auprès des entreprises

Mojca Bavdaz<sup>1</sup>

## Résumé

La connaissance des causes des erreurs de mesure dans les enquêtes auprès des entreprises est limitée, même si ces erreurs peuvent compromettre l'exactitude des microdonnées et des indicateurs économiques qui en découlent. Le présent article, qui est fondé sur une étude empirique axée sur le point de vue des entreprises, présente de nouveaux résultats de recherche sur le processus de réponse aux enquêtes auprès des entreprises. Il propose le modèle multidimensionnel intégral de réponse aux enquêtes auprès des entreprises (MMIRE) comme outil pour examiner le processus de réponse et expliquer ses résultats, et comme base d'une stratégie visant à réduire et à prévenir les erreurs de mesure.

Mots clés : Exactitude ; collecte de données ; statistiques économiques ; enquête auprès des entreprises ; erreur de mesure.

## 1. Introduction

Les erreurs de mesure représentent l'écart entre une mesure idéale et la réponse obtenue à une enquête (Groves, Fowler, Couper, Lepkowski, Singer et Tourangeau 2004). Afin de prévenir efficacement ou de réduire les cas d'erreur de mesure, il est nécessaire de savoir comment le processus de réponse aux questions des enquêtes se déroule et qu'est-ce qui influence son déroulement. Étant donné que les travaux visant à réduire les erreurs dans les enquêtes auprès des entreprises ont traditionnellement été axés sur l'échantillonnage, la base de sondage et les erreurs liées à la non-réponse et, à un degré moindre, sur les erreurs de mesure (Willimack, Lyberg, Martin, Japac et Whitridge 2004), la connaissance des erreurs de mesure et des relations causales sous-jacentes est encore largement limitée dans les enquêtes auprès des entreprises. Le présent article vise à combler cette lacune.

La plupart des études qui examinent les causes des erreurs de mesure dans les enquêtes auprès des entreprises sont un produit de la recherche par prétest. Par conséquent, la majeure partie de ces études sont hypothétiques (par exemple, Morrison, Stettler et Anderson 2002) ou expérimentales (par exemple, Phipps, Butani et Chun 1995), plutôt que d'être fondées sur la collecte proprement dite de données (par exemple, Hak, Willimack et Anderson 2003). L'abondance des résultats de prétest, qui sont habituellement liés à une enquête particulière, contraste avec la rareté des recherches d'évaluation de la qualité (par exemple, Gjesen et Hak 2005), ainsi qu'avec le peu de généralisation et de liens avec le processus de réponse. De nombreuses études sont axées sur un aspect particulier du processus de réponse. Par exemple, Ponikowski et Meily (1989) ont examiné la disponibilité des données nécessaires dans le cadre des

enquêtes auprès des entreprises ; Ramirez (1996) a étudié la sélection des répondants dans les enquêtes auprès des entreprises ; Jenkins et Dillman (1997) ont mené des travaux sur la conception des questionnaires destinés aux entreprises ; O'Brien (2000) et Willimack (2007) ont exploré le rôle des répondants dans la réponse aux enquêtes auprès des établissements ; Greenia, Lane et Willimack (2001) se sont concentrés sur les perceptions qu'ont les entreprises de la confidentialité et sur la question étroitement liée de la mise en commun des données dans les organismes statistiques ; et Willimack (2003) a exposé des problèmes de compréhension et à la mise à l'essai des questionnaires électroniques destinés aux entreprises (par exemple, Snijders, Onat et Visschers 2007), ainsi qu'à leur contrôle (par exemple, Nichols, Murphy, Anderson, Willimack et Sigman 2005), tandis que les plaintes plus fréquentes concernant les coûts imposés par les enquêtes statistiques à la collectivité des entreprises ont suscité des recherches sur le fardeau de réponse.

La première étude portant systématiquement sur l'en- semble du processus de réponse dans les enquêtes auprès des établissements a servi de modèle général au processus de réponse aux enquêtes pour la collecte de données factuelles, qui a été présentée par Edwards et Cantor (1991). Biemer et Fecso (1995) ont combiné le modèle cognitif d'Edwards et Cantor (1991) de réponse aux enquêtes à un modèle statistique qui tentait de quantifier les erreurs de mesure selon leurs sources. Une autre tentative a été faite en 1998-1999, en vue de saisir l'ensemble du processus de réponse dans les enquêtes auprès des entreprises, au moment où le US Census Bureau a mené des interviews qualitatives non structurées sur la production de rapports statistiques. L'étude a servi de base pour deux modèles de réponse aux





- Royall, R.M., et Herson, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.
- Sämdal, C.-E. (1980). On  $\pi$  inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- Sämdal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag, Inc.

- Tillé, Y. (2006). *Sampling Algorithms*. New York : Springer Science+ Business Media, Inc.
- Tillé, Y., et Matei, A. (2005). The R package Sampling. *The Comprehensive R Archive Network*, <http://cran.r-project.org/Manual of the Contributed Packages>.
- Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York : John Wiley & Sons, Inc.

Ensuite, appliquons que  $\phi = 1 - p_i$  pour obtenir l'approximation asymptotique de variance pour la partie linéarisée de  $\bar{y}^{reg}$ .

$$AV(\bar{y}^{reg}) = \sum_{i \in U} (1 - p_i)^3 p_i^3 z_i^3 e_i^2 z_i' M_{zz}^{-1} \bar{z}_N.$$

Nous obtenons l'estimateur de variance en remplaçant les totaux de population par les estimateurs HT sous échantillonnage de Poisson et en intégrant une correction du nombre de degrés de liberté devant  $n(n - s)$  à cause de la petite taille d'échantillon.

### Bibliographie

Beaumont, J.-F., et Bocci, C. (2008). Another look at ridge calibration. *Neuron*, 66, 1, 5-20.

Chambers, R.L. (1996). Robust Case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Fuller, W.A. (1981). An empirical Study of the ratio estimator and estimators of its variance: Comment. *Journal of the American Statistical Association*, 76, 78-80.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.

Fuller, W.A. (2009a). Some design properties of a rejective sampling procedure. *À venir Biometrika*.

Fuller, W.A. (2009b). *Sampling Statistics*. Hoboken, New Jersey : John Wiley & Sons, Inc.

Ikasi, C.T., et Fuller, W.A. (1982). Survey design under a regression population model. *Journal of the American Statistical Association*, 77, 89-96.

Matei, A., et Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21, 543-570.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Rao, J.N.K., et Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, D.C., 57-65.

Rousseau, S., et Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Rapport technique, INSEE, Paris.

Royal, R.M., et Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance – An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

### Annexe

Partons de

$$V(\bar{y}^{reg} | F^v) = V(\bar{y}^{reg} - \bar{y}^v | F^v).$$

Possons que

$$\bar{y}^v = \bar{z}_N' \beta_N$$

et notons que

$$y_i = z_i' \beta_N + e_{Ni}$$

$$\hat{\beta} = \left[ \sum_{i \in A} z_i \phi_i p_i^{-2} z_i' \right]^{-1} \sum_{i \in A} z_i \phi_i p_i^{-2} (z_i' \beta_N + e_i)$$

$$\hat{\beta} = \beta_N + \left[ N^{-1} \sum_{i \in A} z_i \phi_i p_i^{-2} z_i' \right]^{-1} N^{-1} \sum_{i \in A} z_i \phi_i p_i^{-2} e_i. \quad (13)$$

Sous les hypothèses (hypothèses classiques de convergence sous le plan de sondage)

$$N^{-1} \sum_{i \in A} z_i \phi_i p_i^{-2} z_i' = N^{-1} \sum_{i \in U} z_i \phi_i p_i^{-1} z_i' + O_p(n^{-1/2}).$$

Ecrivons

$$N^{-1} \sum_{i \in U} z_i \phi_i p_i^{-1} z_i' = M_{zz,N}.$$

Utilisons le même argument pour développer le terme  $N^{-1} \sum_{i \in A} z_i \phi_i p_i^{-2} e_i$ . Alors, le développement de (13) est

$$\hat{\beta} = \beta_N + M_{zz,N}^{-1} N^{-1} \sum_{i \in A} z_i \phi_i p_i^{-2} e_i + O_p(n^{-1}).$$

Pour construire les intervalles de confiance pour  $\bar{y}^v$ , il est suffisant de prendre en considération la variance du terme linéarisé. Par conséquent, considérons, en utilisant la notation de Särndal, Swensson et Wretman (1992),

$$AV(\bar{y}^{reg}) = \bar{z}_N' M_{zz,N}^{-1} V(\bar{b}^{HT} | F_N) M_{zz,N}^{-1} \bar{z}_N$$

où

$$b_i = z_i \phi_i p_i^{-1} e_i.$$

La variance de l'estimateur HT pour la moyenne de  $b_i$  sous échantillonnage de Poisson est

$$\sum_{i \in U} (1 - p_i) p_i^{-1} b_i b_i'$$



simulations supplémentaires en utilisant des tailles d'échantillon prévues plus grandes ont donné des variances relatives similaires. L'efficacité de l'estimateur par la régression était à peu près la même pour les deux méthodes. Les estimateurs de Horvitz-Thompson calculés en utilisant les probabilités d'inclusion du plan de sondage initial pour ces deux méthodes de correction avaient d'un peu meilleures propriétés que l'estimateur de Horvitz-Thompson obtenu pour la méthode réfective sans contrôle supplémentaire de la taille de l'échantillon.

6. Discussion

L'échantillonnage réfectif et l'échantillonnage par la méthode du cube produisent des estimateurs par la régression dont la performance est à peu près la même. L'échantillonnage produit des gains importants quand le plan de sondage initial donne peu de contrôle sur les valeurs auxiliaires qui entrent dans les échantillons. Un plan d'échantillonnage bien stratifié offre nombre des avantages de l'échantillonnage sur une variable continue. Toutefois, la poursuite de l'échantillonnage après la stratification peut encore donner lieu à de petites améliorations de l'erreur quadratique moyenne pour les estimateurs par la régression. En outre, l'échantillonnage pourrait être utilisé pour prévenir les poids négatifs que produisent les estimateurs par la régression (Fuller 2009a).

Pour les simulations, nous avons fixé le taux de rejet à 90 % pour la population la plus grande. Quand les tailles de la population et de l'échantillon augmentent, il est possible d'accroître le taux de rejet tout en maintenant un grand ensemble d'échantillons possibles. Des simulations supplémentaires ont été exécutées avec des taux de rejet proches de 99 %, mais les données ne sont pas présentées car les différences entre les résultats obtenus avec un taux de 95 % et un taux de 99 % étaient très faibles et le biais de  $\bar{y}^{reg}$  demeurait négligeable. La légère réduction de la variance due à l'échantillonnage diminue à mesure que les conditions d'échantillonnage sont rendues plus rigoureuses.

Dans certains cas particulier, un chercheur pourrait souhaiter effectuer un échantillonnage strict sur certaines variables et plus lâche sur d'autres. Des améliorations peuvent être obtenues en choisissant des poids différents pour les diverses variables ou en répartissant les variables entre des ensembles de test distincts. Les méthodes d'échantillonnage réfectif pondéré et en deux étapes donnent des résultats comparables, si bien que le choix entre ces méthodes dépendra en grande partie de la facilité de mise en œuvre.

Remerciements

Les présents travaux ont été financés aux termes de l'accord de coopération n° 68-3A75-4-122 conclu entre le Natural Resources Conservation Service du USDA et le

La moyenne et la variance en grand échantillon de l'estimateur par la régression sous l'échantillonnage réfectif en deux étapes sont les mêmes que celles de l'estimateur par la régression sous le plan de sondage original. En outre, l'estimateur habituel de la variance sous le plan de sondage original pour l'estimateur par la régression convient pour l'échantillonnage réfectif en deux étapes. La preuve de cette déclaration, qui est une extension de la preuve donnée dans Fuller (2009a), peut être obtenue sur demande.

Afin d'examiner certaines propriétés des deux méthodes, nous avons répété les simulations Monte Carlo pour le plan d'échantillonnage de Poisson initial avec la variable  $p_i$  séparée des trois autres variables. Nous avons transformé le vecteur de variables d'échantillonnage de façon que la matrice de variance des estimateurs de Horvitz-Thompson des totaux soit diagonale. Pour la méthode avec pondération, nous avons fixé le poids appliqué à la composante  $p_i$  de forme quadratique à 1,5, ceux appliqués aux autres composantes, à 1 et nous avons donné à  $\gamma$  la valeur 0,627. Cette méthode de pondération limitait les échantillons à ceux dont la taille variait de 18 à 22. Pour la méthode en deux étapes, tout échantillon dont la taille se situait en dehors de la fourchette de 18 à 22 a été rejeté à la première étape, puis la forme quadratique pour les trois variables restantes a été vérifiée en utilisant une valeur de  $\gamma$  de 0,63 pour la deuxième étape. Étant donné les bonnes propriétés de l'estimateur de variance  $V(\bar{y}^{reg})$  donné par (10), le tableau 4 ne contient que ses valeurs Monte Carlo moyennes  $\text{moy}(V(\bar{y}^{reg}))$ .

Tableau 4 Propriétés des échantillons de taille prévue de 20 obtenus par la méthode réfective avec corrections basées sur l'échantillonnage de Poisson, et taux de rejet de 95 %

Pondération		Deux étapes	
$\text{biais}_\pi(\bar{y}^{HT}) / \sqrt{V^d(\bar{y}^{HT})}$	-0,005	-0,014	
$\text{biais}_\pi(\bar{y}^{reg}) / \sqrt{V^d(\bar{y}^{HT})}$	0,003	0,002	
$V(\bar{y}^{HT}) / A^d(\bar{y}^{HT})$	0,210	0,217	
$V(\bar{y}^{reg}) / A^d(\bar{y}^{HT})$	0,132	0,132	
$\text{moy}(V(\bar{y}^{reg})) / V^d(\bar{y}^{HT})$	0,121	0,121	
$V^\pi(n)$	1,237	1,902	

Les résultats obtenus pour une taille d'échantillon prévue de 20 et un taux de rejet proche de 95 % étaient semblables pour les deux méthodes de correction (tableau 4). L'estimateur de Horvitz-Thompson calculé pour la méthode pondérée donnait d'un peu meilleurs résultats que celui calculé pour la méthode en deux étapes. L'une des raisons de cette différence est que la variation des tailles d'échantillon était nettement plus faible dans le cas de la méthode pondérée ( $V^\pi(n)$ ) à la dernière ligne du tableau 4). Des

deuxième approche nécessite certaines hypothèses de plus que celles de Fuller (2009a), mais un argument similaire peut être utilisé pour justifier la procédure.

Afin de prouver les propriétés de convergence de la méthode de rejet à tests multiples, il est commode de considérer deux sous-ensembles de variables d'équilibrage et d'imaginer que le rejet est effectué séquentiellement sur chaque sous-ensemble. Nous donnons à la méthode de rejet sur deux sous-ensembles le nom de méthode d'échantillonnage réjectif en deux étapes. Supposons que  $z'_i = (z'_{i1}, z'_{i2})$  est le vecteur de variables d'équilibrage et que le plan de sondage original est désigné par  $p(\cdot)$ . La procédure est la suivante.

Étape 1 : Sélectionner un échantillon en utilisant  $p(\cdot)$  et rejeter les échantillons en appliquant la condition d'équilibrage (8) au premier sous-ensemble  $z_{i1}$ .

$$\tilde{O}_1 = (\bar{z}_{HT,1} - \bar{z}_{N,1})' A(\bar{z}_{HT,1} | F^N)^{-1} (\bar{z}_{HT,1} - \bar{z}_{N,1}) > \gamma_1.$$

Étape 2 : Utiliser l'échantillon accepté à l'étape 1 pour vérifier la condition d'équilibrage (8) sur le deuxième sous-

échantillon  $z_{i2}$ .

$$\tilde{O}_2 = (\bar{z}_{HT,2} - \bar{z}_{N,2})' A(\bar{z}_{HT,2} | F^N)^{-1} (\bar{z}_{HT,2} - \bar{z}_{N,2}) > \gamma_2.$$

Rejeter l'échantillon si la condition n'est pas satisfaite et répéter l'étape 1.

En pratique, aussi bien pour l'échantillonnage réjectif

pondéré que pour l'échantillonnage réjectif en deux étapes, des ratonnements sont vraisemblablement nécessaires pour choisir les valeurs de  $\gamma$ . Dans le cas de la méthode pondérée, la forme quadratique devient une somme de multiples des variables aléatoires  $\chi^2$ , ce qui rend le choix de  $\gamma$  plus difficile que dans le cas non pondéré. Nous avons utilisé des approximations par appariement des moments pour choisir les valeurs de  $\gamma$  qui fournissent des taux de rejet proches de ceux souhaités, mais nous avons ensuite procédé à de petites simulations pour déterminer le taux de rejet sous forme d'une fonction de  $\gamma$ . Pour la méthode en deux étapes, nous avons utilisé une approximation par une loi du  $\chi^2$  pour sélectionner une valeur de  $\gamma_1$  donnant approximativement le taux de rejet souhaité à la première étape et nous avons utilisé une deuxième approximation par le  $\chi^2$  pour sélectionner une valeur initiale de  $\gamma_2$  donnant approximativement le taux de rejet souhaité à la deuxième étape. Nous avons ajusté le deuxième paramètre  $\gamma_2$  afin d'obtenir le taux de rejet global cible. Le choix des valeurs de  $\gamma$  dans la méthode en deux étapes est subjectif, car de nombreuses combinaisons de  $\gamma_1$  et  $\gamma_2$  peuvent produire le même taux global. En pratique, un praticien imposera vraisemblablement un bonrage strict pour le premier sous-ensemble de variables et des bornes lâches sur les variables d'équilibrage restantes.

constant. Il est possible de modifier la méthode d'échantillonnage réjectif de manière que l'équilibre sur  $p_i$  soit plus strict que sur d'autres variables.

Un moyen d'accroître l'équilibrage sur un sous-ensemble de variables consiste à modifier la fonction de test de rejet. Dans Fuller (2009a), l'ordre de l'approximation des probabilités d'inclusion de premier et de deuxième ordres demeure le même quand, dans la forme quadratique de rejet, la matrice de variance est remplacée par une matrice définie positive symétrique de même ordre.

Pour déterminer les pondérations pour l'échantillonnage réjectif pondéré, il est commode de transformer les variables d'équilibrage de façon que  $A(\bar{z}_{HT} | F^N)$  soit une matrice diagonale. Pour l'échantillonnage réjectif pondéré, la statistique de test est

$$(12) \quad \sum_{i=1}^q c_i^q A(\bar{z}_{HT,q} | F^N)^{-1} (\bar{z}_{HT,q} - \bar{z}_{N,q})^2,$$

où  $m$  est le nombre de variables d'équilibrage,  $z_q$  est la  $q^{\text{e}}$  variable d'équilibrage et  $c_q$  représente les poids sélectionnés. Nous pouvons donner au poids appliqué à la première variable  $z_{i1} = p_i$  une grande valeur relativement à celle des poids appliqués aux autres variables afin de réduire la variance de la taille de l'échantillon. La transformation utilisée est celle de Gramm-Schmidt en se servant de la variance sous le plan de sondage initial. L'équilibrage est effectué sur les variables transformées, mais la première variable n'est pas transformée. Les estimateurs de Horvitz-Thompson des variables transformées ne sont pas corrélés. L'équilibrage sur les variables transformées produira encore un équilibre sur les variables originales, puisque chaque variable transformée est le résidu d'une opération de régression sur les variables précédentes.

Un parallèle peut être établi entre l'équation (12) et le terme de pénalité de la fonction de distance qui sous-tend le calage par la régression Ridge. Voir Rao et Singh (1997), Beaumont et Bocci (2008), ainsi que Chambers (1996). En particulier, le choix des poids  $c_q$  est semblable au problème de sélection des coûts appropriés dans le calage par la régression Ridge. Donc, l'échantillonnage réjectif en utilisant la statistique de test (12) peut être considéré comme l'intégration du calage par la régression Ridge à l'étape de l'élaboration du plan de sondage.

Un deuxième moyen de produire un équilibre plus strict sur un sous-ensemble de variables consiste à effectuer le rejet séparément pour divers sous-ensembles. Une statistique de test est produite pour chaque sous-ensemble et, pour être acceptée, il faut qu'un échantillon soit accepté à tous les tests. Dans le cas de l'échantillonnage de Poisson, d'échantillon n'est pas comprise entre les limites de tolérance spécifiées pour la taille d'échantillon prévue. Cette



à l'estimateur de Horvitz-Thompson calculé en utilisant les probabilités d'inclusion du plan initial. Le biais et la variance de  $\hat{y}_{HT}$  sont proches de ceux de  $\hat{y}_{reg}$  sous la méthode du cube ainsi que la méthode réjective. Le biais estimé plus grand de  $\hat{y}_{HT}$  sous l'échantillonnage par la méthode du cube est dû à l'erreur de Monte Carlo. L'amélioration due à l'équilibrage sur  $x$  n'est pas importante comparativement à celle obtenue pour l'exemple d'échantillonnage de Poisson. Cependant, sous ce plan stratifié initial fortement contrôlé, dans lequel les échantillons initiaux sont déjà partiellement équilibrés sur  $x$ , un équilibrage supplémentaire et l'utilisation des estimateurs  $\hat{y}_{reg}$  peuvent encore offrir un avantage modéré. Ce résultat peut être constaté pour  $\hat{y}_{reg}$  en comparant la quatrième ligne du tableau 2 à la variance de  $\hat{y}_{reg}$  sous le plan initial  $V^d(\hat{y}_{reg}) = 0,987$ . Par conséquent, dans ce cas, une bonne stratégie consiste à combiner la stratification, l'équilibrage et la régression, conclusion qui est semblable à celle tirée par Deville et Tillé (2004). L'estimateur de variance  $V^d(\hat{y}_{reg})$  obtenu en utilisant (11) donne des estimations qui, en moyenne, pour les variances de l'estimateur par la régression sous la méthode du cube ainsi que la méthode réjective, sont proches des variances réelles. Cependant, l'estimateur de variance  $V^{DT}(\hat{y}_{reg})$  proposé par Deville et Tillé (2005) donne de l'échantillonnage de Poisson dans le deuxième estimateur de variance de Deville et Tillé (2005) repose sur l'hypothèse que les probabilités d'inclusion conjointe sont éloignées des probabilités d'inclusion conjointe réelles dans les petites strates. Or, les probabilités d'inclusion conjointe dans les petites strates sont plus proches de celles de l'échantillonnage aléatoire stratifié que de celles de l'échantillonnage de Poisson. Ce problème pourrait expliquer pourquoi  $V^d(\hat{y}_{reg})$  donné par (11) en utilisant les probabilités initiales d'inclusion pour deux unités par strate est moins biaisé que  $V^{DT}(\hat{y}_{reg})$  dans ce cas.

**Tableau 2**  
Propriétés des échantillons de taille prévue de 20 basés sur l'échantillonnage stratifié.  $V^d(\hat{y}_{HT}) = 0,011$  et  $V^d(\hat{y}_{reg})/V^d(\hat{y}_{HT}) = 0,987$

	Cube	Réf.	90 %	95 %
$\text{biais}_{\pi}(\hat{y}_{HT}) / \sqrt{V^d(\hat{y}_{HT})}$	-0,028	0,014	0,010	0,010
$\text{biais}_{\pi}(\hat{y}_{reg}) / \sqrt{V^d(\hat{y}_{HT})}$	-0,013	0,014	0,010	0,010
$V^{\pi}(\hat{y}_{HT}) / V^d(\hat{y}_{HT})$	0,910	0,866	0,813	0,813
$V^{\pi}(\hat{y}_{reg}) / V^d(\hat{y}_{HT})$	0,929	0,865	0,813	0,813
$\text{moy}(V^d(\hat{y}_{reg})) / V^d(\hat{y}_{HT})$	0,907	0,881	0,775	0,775
$\text{moy}(V^{DT}(\hat{y}_{reg})) / V^d(\hat{y}_{HT})$	0,792	-	-	-

Afin d'évaluer les propriétés en grand échantillon des méthodes d'équilibrage, nous avons quadruplé la taille de la simulation de Poisson. Nous avons répété quatre fois la population et sélectionné un échantillon de taille prévue de 80. La variance d'une moyenne de Horvitz-Thompson sous l'échantillonnage de Poisson est  $V^d(\hat{y}_{HT}) = 0,020$  et la variance de l'estimateur par la régression est  $V^d(\hat{y}_{reg}) = 0,132$ . Les variances et biais relatifs résultants sont proches des résultats pour les échantillons de taille 20 (tableau 3). Les résultats des simulations corroborent le résultat théorique de Fuller (2009a) selon lequel l'estimateur par la régression est un estimateur d'ordre  $O_p(n^{-1/2})$  après un rejet du type utilisé dans le présent article. Bien que cela ne soit pas prouvé ici, l'estimateur par la régression utilisé après l'échantillonnage par la méthode du cube semble posséder des propriétés semblables à l'estimateur par la régression lorsque l'on utilise l'échantillonnage réjectif.

**Tableau 3**  
Propriétés des échantillons de taille prévue de 80 basés sur l'échantillonnage de Poisson.  $V^d(\hat{y}_{HT}) = 0,02$  et  $V^d(\hat{y}_{reg})/V^d(\hat{y}_{HT}) = 0,132$

	Cube	Réf.	90 %	95 %
$\text{biais}_{\pi}(\hat{y}_{HT}) / \sqrt{V^d(\hat{y}_{HT})}$	0,002	-0,006	-0,007	-0,007
$\text{biais}_{\pi}(\hat{y}_{reg}) / \sqrt{V^d(\hat{y}_{HT})}$	0,002	0,000	-0,001	-0,001
$V^{\pi}(\hat{y}_{HT}) / V^d(\hat{y}_{HT})$	0,127	0,267	0,224	0,224
$V^{\pi}(\hat{y}_{reg}) / V^d(\hat{y}_{HT})$	0,122	0,124	0,123	0,123
$\text{moy}(V^d(\hat{y}_{reg})) / V^d(\hat{y}_{HT})$	0,121	0,121	0,121	0,121
$\text{moy}(V^{DT}(\hat{y}_{reg})) / V^d(\hat{y}_{HT})$	0,121	-	-	-

## 5. Corrections apportées à la méthode de rejet

Dans la méthode d'échantillonnage réjectif de Fuller, la même importance est accordée à toutes les variables d'équilibrage. Pour un grand nombre de celles-ci, on ne peut s'attendre à un équilibre exact sur toutes les variables et l'approximation pourrait être médiocre pour certaines variables importantes. Par conséquent, un praticien pourrait souhaiter obtenir un équilibre plus strict sur un sous-ensemble de variables d'équilibrage. Par exemple, un chercheur pourrait vouloir utiliser l'échantillonnage de Poisson par souci de simplicité, mais exercer aussi un certain contrôle sur la taille aléatoire de l'échantillon. Une taille d'échantillon aléatoire peut compliquer la planification de l'étude et contribuer fortement à la variance des estimateurs. L'échantillonnage équilibré permet de réduire la variation des tailles d'échantillon en équilibrant sur la variable  $p_i$ , qui est la probabilité d'inclusion de premier ordre initiale. Dans la méthode d'échantillonnage réjectif de Fuller, la variance de la taille d'échantillon s'accroît quand le nombre de variables d'équilibrage augmente et que le taux de rejet est maintenu



erreurs quadratiques moyennes sont réduites encore davantage en utilisant l'estimateur par la régression avec l'une ou l'autre méthode d'équilibrage. Le gain dû à l'utilisation de l'estimateur par la régression est plus important pour l'échantillonnage réjectif que pour l'échantillonnage par la méthode du cube, vraisemblablement parce que cette dernière produit un équilibre plus strict que la méthode réjective. Les deux méthodes donnent lieu à des variances similaires pour l'estimateur par la régression. La variance de l'estimateur par la régression sous le plan initial de Poisson est  $V^p(\bar{y}^{reg}) = 0,249$  (relativement à  $V^p(\bar{y}^{HT})$ ). En comparant la valeur de 0,249 à la quatrième ligne du tableau 1, nous voyons que, pour l'estimateur par la régression, le gain résultant de l'utilisation d'échantillons équilibrés est modéré. Ce résultat est en harmonie avec la constatation de Fuller (2009a) selon laquelle la réduction de la variance de  $\bar{y}^{reg}$  en utilisant des échantillons obtenus par la méthode réjective est due à une correction de deuxième ordre. L'estimateur de variance de  $\bar{y}^{reg}$  en utilisant (10) présente un petit biais tant pour les échantillons obtenus par la méthode du cube que pour ceux produits par la méthode réjective ( $\text{moy}(V^p(\bar{y}^{reg}))$  dans le tableau 1). L'estimateur de variance  $V^{DT}(\bar{y}^{reg})$  posé dans Deville et Tillé (2005) donne des résultats comparables à  $V^p(\bar{y}^{reg})$  dans (10) puisque le deuxième estimateur de variance de Deville et Tillé (2005) est très proche de (10) pour l'échantillonnage de Poisson. Ce résultat appuie l'allégation selon laquelle, dans les estimateurs de variance de Deville et Tillé (2005), l'hypothèse de l'approximation de Poisson est satisfaite pour le cas du plan d'échantillonnage de Poisson.

**Tableau 1**  
Propriétés des échantillons de taille prévue de 20 basés sur l'échantillonnage de Poisson.  $V^p(\bar{y}^{HT}) = 0,08$  et  $V^p(\bar{y}^{reg})/V^p(\bar{y}^{HT}) = 0,249$

Cube	Réf. 90 %	Réf. 95 %
biases $\pi(\bar{y}^{HT}) / \sqrt{V^p(\bar{y}^{HT})}$	-0,002	-0,016
biases $\pi(\bar{y}^{reg}) / \sqrt{V^p(\bar{y}^{HT})}$	-0,002	0,005
$V^p(\bar{y}^{HT})/V^p(\bar{y}^{HT})$	0,142	0,220
$V^p(\bar{y}^{reg})/V^p(\bar{y}^{HT})$	0,131	0,129
$\text{moy}(V^p(\bar{y}^{reg}))/V^p(\bar{y}^{HT})$	0,122	0,123
$\text{moy}(V^{DT}(\bar{y}^{reg}))/V^p(\bar{y}^{HT})$	0,120	-

Au tableau 2, nous présentons les estimations sous le plan d'échantillonnage stratifié à deux unités par strate initial. La variance de la moyenne de Horvitz-Thompson sous le plan de stratification initial est  $V^p(\bar{y}^{HT}) = 0,011$  et toutes les estimations sont normalisées au moyen de cette valeur. Puisque, dans ce plan initial, la stratification contrôle la plupart de l'effet de  $x$  sur  $y$ , l'estimateur par la régression n'offre pas d'amélioration importante par rapport

de Poisson avec des corrections pour les contraintes connues décrits dans Deville et Tillé (2005) ne différant que légèrement, nous n'avons utilisé que le deuxième dans les études par simulation. Deville et Tillé (2005) proposent aussi un quatrième estimateur, mais celui-ci requiert la résolution d'un système d'équations non linéaires dont l'ajout à la simulation aurait demandé beaucoup de ressources informatiques. Cependant, leur quatrième estimateur pourrait donner de meilleurs résultats que les autres pour les plans d'échantillonnage stratifiés, puisqu'il reproduit la variance d'un échantillon aléatoire stratifié quand le vecteur de variables d'équilibrage contient des indicateurs de strate.

- $V^p(\bar{y}^{reg})$  : variance estimée de l'estimateur par la régression en utilisant l'équation (10) (ou (11)) pour le plan de sondage initial de Poisson (ou stratifié à deux unités par strate) et chaque échantillon équilibré.
- $\text{moy}(V^p(\bar{y}^{reg}))$  : moyenne Monte Carlo de  $V^p(\bar{y}^{reg})$  en utilisant les échantillons équilibrés.

Dans les simulations, nous avons également calculé  $V^p(\bar{y}^{reg})$  pour les échantillons obtenus par la méthode du cube aux fins de comparaison.

Le tableau 1 donne les estimations pour le plan d'échantillonnage de Poisson. La variance de la moyenne de Horvitz-Thompson sous l'échantillonnage de Poisson initial avec taille de l'échantillon prévue de 20 et sans équilibrage est  $V^p(\bar{y}^{HT}) = 0,08$ . Dans le tableau 1, les variances sont normalisées par  $V^p(\bar{y}^{HT})$ , et les biais sont normalisés par  $\sqrt{V^p(\bar{y}^{HT})}$ . L'estimateur de Horvitz-Thompson est sans biais sous les plans obtenus par la méthode du cube, parce que l'échantillonnage par la méthode du cube retient les probabilités d'inclusion de premier ordre. L'estimateur de Horvitz-Thompson en utilisant les probabilités d'inclusion du plan initial est biaisé sous l'échantillonnage réjectif, parce que les probabilités d'inclusion diffèrent des probabilités d'inclusion du plan initial, comme l'indique la figure 1. Le biais de l'estimateur par la régression sous échantillonnage réjectif est plus faible que celui de l'estimateur de Horvitz-Thompson calculé avec les probabilités d'inclusion du plan initial. Le biais de  $\bar{y}^{reg}$  est du même ordre sous la méthode du cube et la méthode réjective. L'accroissement du taux de rejet augmente le biais de  $\bar{y}^{reg}$  pour les plans obtenus par la méthode réjective. Cependant, sous les deux méthodes d'équilibrage et les deux taux de rejet, les biais de  $\bar{y}^{reg}$  sont négligeables comparativement aux variances de Monte Carlo. Pour l'estimateur de Horvitz-Thompson en utilisant les probabilités d'inclusion du plan initial, le gain réalisé en utilisant l'échantillon équilibré est important pour la méthode du cube ainsi que la méthode réjective. Les

Afin d'évaluer les propriétés en grand échantillon des méthodes d'équilibrage, nous avons quadruplé la taille de la simulation de Poisson. Nous avons répété quatre fois la population et sélectionné un échantillon de taille prévue de 80. La variance d'une moyenne de Horvitz-Thompson sous l'échantillonnage de Poisson est  $V^d(\bar{Y}_{HT}) = 0,020$  et la variance de l'estimateur par la régression est  $V^d(\bar{Y}^{reg}) = 0,132$ . Les variances et biais relatifs résultants sont proches des résultats pour les échantillons de taille 20 (tableau 3). Les résultats des simulations corroborent le résultat théorique de Fuller (2009a) selon lequel l'estimateur par la régression est un estimateur d'ordre  $O_p(n^{-1/2})$  après un rejet du type utile utilisé dans le présent article. Bien que cela ne soit pas prouvé ici, l'estimateur par la régression utilisé après l'échantillonnage par la méthode du cube semble posséder des propriétés semblables à l'estimateur par la régression lorsque l'on utilise l'échantillonnage réjectif.

**Tableau 3**  
Propriétés des échantillons de taille prévue de 80 basés sur l'échantillonnage de Poisson.  $V^d(\bar{Y}_{HT}) = 0,02$  et  $V^d(\bar{Y}^{reg}) / V^d(\bar{Y}_{HT}) = 0,132$

Cube		Réf.		Réf.	
90 %		95 %		95 %	
biais $\pi(\bar{Y}_{HT}) / \sqrt{V^d(\bar{Y}_{HT})}$	0,002	-0,006	-0,007	biais $\pi(\bar{Y}^{reg}) / \sqrt{V^d(\bar{Y}^{reg})}$	-0,001
$V^{\pi}(\bar{Y}_{HT}) / V^d(\bar{Y}_{HT})$	0,127	0,267	0,224	$V^{\pi}(\bar{Y}^{reg}) / V^d(\bar{Y}^{reg})$	0,123
$V^{\pi}(\bar{Y}^{reg}) / V^d(\bar{Y}_{HT})$	0,122	0,124	0,121	$\text{moy}(\bar{Y}^{reg}) / V^d(\bar{Y}_{HT})$	0,121
$\text{moy}(\bar{Y}^{reg}) / V^d(\bar{Y}_{HT})$	0,121	-	-	$\text{moy}(\bar{Y}_{HT}) / V^d(\bar{Y}^{reg})$	-

## 5. Corrections apportées à la méthode de rejet

Dans la méthode d'échantillonnage réjectif de Fuller, la même importance est accordée à toutes les variables d'équilibrage. Pour un grand nombre de celles-ci, on ne peut s'attendre à un équilibre exact sur toutes les variables et l'approximation pourrait être médiocre pour certaines variables importantes. Par conséquent, un praticien pourrait souhaiter obtenir un équilibre plus strict sur un sous-ensemble de variables d'équilibrage. Par exemple, un chercheur pourrait vouloir utiliser l'échantillonnage de Poisson par souci de simplicité, mais exercer aussi un certain contrôle sur la taille aléatoire de l'échantillon. Une taille d'échantillon aléatoire peut compliquer la planification de l'étude et contribuer fortement à la variance des estimateurs. L'échantillonnage équilibré permet de réduire la variation des tailles d'échantillon en équilibrant sur la variable  $p_i$ , qui est la probabilité d'inclusion de premier ordre initiale. Dans la méthode d'échantillonnage réjectif de Fuller, la variance de la taille d'échantillon s'accroît quand le nombre de variables d'équilibrage augmente et que le taux de rejet est maintenu

à l'estimateur de Horvitz-Thompson calculé en utilisant les probabilités d'inclusion du plan initial. Le biais et la variance de  $\bar{Y}_{HT}$  sont proches de ceux de  $\bar{Y}^{reg}$  sous la méthode du cube ainsi que la méthode réjective. Le biais estimé plus grand de  $\bar{Y}_{HT}$  sous l'échantillonnage par la méthode du cube est dû à l'erreur de Monte Carlo. L'amélioration due à l'équilibrage sur  $x$  n'est pas importante comparativement à celle obtenue pour l'exemple d'échantillonnage de Poisson. Cependant, sous ce plan stratifié initial fortement contrôlé, dans lequel les échantillons initiaux sont déjà partiellement équilibrés sur  $x$ , un équilibrage supplémentaire et l'utilisation des estimateurs  $\bar{Y}^{reg}$  peuvent encore offrir un avantage modéré. Ce résultat peut être constaté pour  $\bar{Y}^{reg}$  en comparant la quatrième ligne du tableau 2 à la variance de  $\bar{Y}^{reg}$  sous le plan initial  $V^d(\bar{Y}^{reg}) = 0,987$ . Par conséquent, dans ce cas, une bonne stratégie consiste à combiner la stratification, l'équilibrage et la régression, conclusion qui est semblable à celle tirée par Deville et Tillé (2004). L'estimateur de variance  $V^d(\bar{Y}^{reg})$  obtenu en utilisant (11) donne des estimations qui, en moyenne, pour les variances de l'estimateur par la régression sous la méthode du cube ainsi que la méthode réjective, sont proches des variances réelles. Cependant, l'estimateur de variance  $V^{DT}(\bar{Y}^{reg})$  proposé par Deville et Tillé (2005) donne de l'échantillonnage de Poisson dans le deuxième estimateur de variance de Deville et Tillé (2005) repose sur l'hypothèse que les probabilités d'inclusion conjointe sont éloignées des probabilités d'inclusion conjointe réelles dans les petites strates. Or, les probabilités d'inclusion conjointe dans les petites strates sont plus proches de celles de l'échantillonnage aléatoire stratifié que de celles de l'échantillonnage de Poisson. Ce problème pourrait expliquer pourquoi  $V^d(\bar{Y}^{reg})$  donné par (11) en utilisant les probabilités initiales d'inclusion pour deux unités par strate est moins biaisé que  $V^{DT}(\bar{Y}^{reg})$  dans ce cas.

**Tableau 2**  
Propriétés des échantillons de taille prévue de 20 basés sur l'échantillonnage stratifié.  $V^d(\bar{Y}_{HT}) = 0,011$  et  $V^d(\bar{Y}^{reg}) / V^d(\bar{Y}_{HT}) = 0,987$

Cube		Réf.		Réf.	
90 %		95 %		95 %	
biais $\pi(\bar{Y}_{HT}) / \sqrt{V^d(\bar{Y}_{HT})}$	-0,028	0,014	0,010	biais $\pi(\bar{Y}^{reg}) / \sqrt{V^d(\bar{Y}_{HT})}$	0,010
$V^{\pi}(\bar{Y}_{HT}) / V^d(\bar{Y}_{HT})$	0,910	0,866	0,813	$V^{\pi}(\bar{Y}^{reg}) / V^d(\bar{Y}_{HT})$	0,813
$V^{\pi}(\bar{Y}^{reg}) / V^d(\bar{Y}_{HT})$	0,929	0,865	0,813	$\text{moy}(\bar{Y}^{reg}) / V^d(\bar{Y}_{HT})$	0,775
$\text{moy}(\bar{Y}^{reg}) / V^d(\bar{Y}_{HT})$	0,907	0,881	-	$\text{moy}(\bar{Y}_{HT}) / V^d(\bar{Y}^{reg})$	-
$\text{moy}(\bar{Y}_{HT}) / V^d(\bar{Y}^{reg})$	0,792	-	-		



erreurs quadratiques moyennes sont réduites encore davantage en utilisant l'estimateur par la régression avec l'une ou l'autre méthode d'équilibrage. Le gain dû à l'utilisation de l'estimateur par la régression est plus important pour l'échantillonnage réjectif que pour l'échantillonnage par la méthode du cube, vraisemblablement parce que cette dernière produit un équilibre plus strict que la méthode réjective. Les deux méthodes donnent lieu à des variances similaires pour l'estimateur par la régression. La variance de l'estimateur par la régression sous le plan initial de Poisson est  $V^p(\bar{y}^{reg}) = 0,249$  (relativement à  $V^p(\bar{y}^{HT})$ ). En comparant la valeur de 0,249 à la quatrième ligne du tableau 1, nous voyons que, pour l'estimateur par la régression, le gain résultant de l'utilisation d'échantillons équilibrés est modéré. Ce résultat est en harmonie avec la constatation de Fuller (2009a) selon laquelle la réduction de la variance de  $\bar{y}^{reg}$  en utilisant des échantillons obtenus par la méthode réjective est due à une correction de deuxième ordre. L'estimateur de variance de  $\bar{y}^{reg}$  en utilisant (10) présente un petit biais tant pour les échantillons obtenus par la méthode du cube que pour ceux produits par la méthode réjective ( $\text{moy}(V^p(\bar{y}^{reg}))$  dans le tableau 1). L'estimateur de variance  $V^{DT}(\bar{y}^{reg})$  posé dans Deville et Tillé (2005) donne des résultats comparables à  $V(\bar{y}^{reg})$  dans (10) puisque le deuxième estimateur de variance de Deville et Tillé (2005) est très proche de (10) pour l'échantillonnage de Poisson. Ce résultat appuie l'allégation selon laquelle, dans les estimateurs de variance de Deville et Tillé (2005), l'hypothèse de l'approximation de Poisson est satisfaite pour le cas du plan d'échantillonnage de Poisson.

**Tableau 1**  
Propriétés des échantillons de taille prévue de 20 basés sur l'échantillonnage de Poisson.  $V^p(\bar{y}^{HT}) = 0,08$  et  $V^p(\bar{y}^{reg})/V^p(\bar{y}^{HT}) = 0,249$

Cube	Réf. 90 %	Réf. 95 %
biais $_{\pi}(\bar{y}^{HT})/\sqrt{V^p(\bar{y}^{HT})}$	-0,002	-0,016
biais $_{\pi}(\bar{y}^{reg})/\sqrt{V^p(\bar{y}^{HT})}$	-0,002	0,002
$V^p(\bar{y}^{HT})/V^p(\bar{y}^{reg})$	0,142	0,270
$V^p(\bar{y}^{reg})/V^p(\bar{y}^{HT})$	0,131	0,136
$\text{moy}(V^p(\bar{y}^{reg}))/V^p(\bar{y}^{HT})$	0,122	0,123
$\text{moy}(V^{DT}(\bar{y}^{reg}))/V^p(\bar{y}^{HT})$	0,120	-

Au tableau 2, nous présentons les estimations sous le plan d'échantillonnage stratifié à deux unités par strate initial. La variance de la moyenne de Horvitz-Thompson sous le plan de stratification initial est  $V^p(\bar{y}^{HT}) = 0,011$  et toutes les estimations sont normalisées au moyen de cette valeur. Puisque, dans ce plan initial, la stratification contrôle la plupart de l'effet de  $x$  sur  $y$ , l'estimateur par la régression n'offre pas d'amélioration importante par rapport

de Poisson avec des corrections pour les contraintes connues décrits dans Deville et Tillé (2005) ne différant que légèrement, nous n'avons utilisé que le deuxième dans les études par simulation. Deville et Tillé (2005) proposent aussi un quatrième estimateur, mais celui-ci requiert la résolution d'un système d'équations non linéaires dont l'ajout à la simulation aurait demandé beaucoup de ressources informatiques. Cependant, leur quatrième estimateur pourrait donner de meilleurs résultats que les autres pour les plans d'échantillonnage stratifiés, puisqu'il reproduit la variance d'un échantillon aléatoire stratifié quand le vecteur de variables d'équilibrage contient des indicateurs de strate.

- $V^p(\bar{y}^{reg})$  : variance estimée de l'estimateur par la régression en utilisant l'équation (10) (ou (11)) pour le plan de sondage initial de Poisson (ou stratifié à deux unités par strate) et chaque échantillon équilibré.
- $\text{moy}(V^p(\bar{y}^{reg}))$  : moyenne Monte Carlo de  $V^p(\bar{y}^{reg})$  en utilisant les échantillons équilibrés.

Dans les simulations, nous avons également calculé  $V^p(\bar{y}^{reg})$  pour les échantillons obtenus par la méthode du cube aux fins de comparaison.

Le tableau 1 donne les estimations pour le plan d'échantillonnage de Poisson. La variance de la moyenne de Horvitz-Thompson sous l'échantillonnage de Poisson initial avec taille de l'échantillon prévue de 20 et sans équilibrage est  $V^p(\bar{y}^{HT}) = 0,08$ . Dans le tableau 1, les variances sont normalisées par  $V^p(\bar{y}^{HT})$ , et les biais sont normalisés par  $\sqrt{V^p(\bar{y}^{HT})}$ . L'estimateur de Horvitz-Thompson est sans biais sous les plans obtenus par la méthode du cube, parce que l'échantillonnage par la méthode du cube retient les probabilités d'inclusion de premier ordre. L'estimateur de Horvitz-Thompson en utilisant les probabilités d'inclusion du plan initial est biaisé sous l'échantillonnage réjectif, parce que les probabilités d'inclusion diffèrent des probabilités d'inclusion du plan initial, comme l'indique la figure 1. Le biais de l'estimateur par la régression sous échantillonnage réjectif est plus faible que celui de l'estimateur de Horvitz-Thompson calculé avec les probabilités d'inclusion du plan initial. Le biais de  $\bar{y}^{reg}$  est du même ordre sous la méthode du cube et la méthode réjective. L'accroissement du taux de rejet augmente le biais de  $\bar{y}^{reg}$  pour les plans obtenus par la méthode réjective. Cependant, sous les deux méthodes d'équilibrage et les deux taux de rejet, les biais de  $\bar{y}^{reg}$  sont négligeables comparativement aux variances de Monte Carlo. Pour l'estimateur de Horvitz-Thompson en utilisant les probabilités d'inclusion du plan initial, le gain réalisé en utilisant l'échantillon équilibré est important pour la méthode du cube ainsi que la méthode réjective. Les





probabilités d'inclusion égales dans la strate. Nous avons choisi les limites des strates de cette façon pour que la probabilité d'inclusion de l'unité  $i$  soit presque proportionnelle à  $x_i$ , ce qui est la probabilité d'inclusion optimale sous le modèle (9) (Ikasi et Fuller 1982). Ce plan de stratification peut aussi être équilibré partiellement sur  $x$  par la voie d'un plan de sondage standard. Dans le plan d'échantillonnage aléatoire stratifié, l'équilibre est atteint en utilisant une fonction escalier pour approximer une droite. Le plan stratifié sera également partiellement équilibré sur  $x^2$ . Le plan d'échantillonnage aléatoire stratifié est destiné à illustrer dans quelle mesure un équilibrage supplémentaire peut être avantageux. Nous avons tiré deux unités par strate afin d'obtenir le nombre maximal de strates tout en permettant une estimation de variance sans biais. Fuller (1981) a montré que, dans le cas de deux unités par strate, ce plan d'échantillonnage stratifié possède une variance anticipée proche de la variance du meilleur modèle d'échantillonnage à choix raisonné sous (4). Pour l'échantillonnage de Poisson, nous avons pris, pour la taille d'échantillon prévue de 20, des probabilités d'inclusion initiales égales aux probabilités d'inclusion initiales du plan d'échantillonnage stratifié.

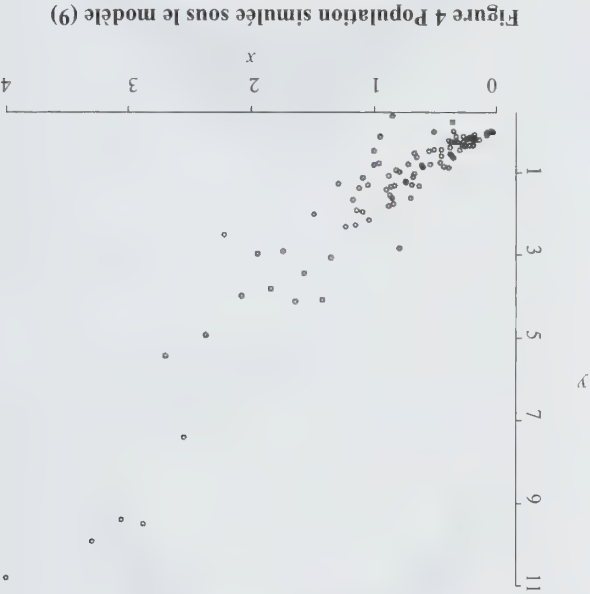


Figure 4 Population simulée sous le modèle (9)

L'estimateur par la régression étudié dans le présent article est de la forme (1) avec  $\beta$  défini en (2). La variable de régression  $z$  est un vecteur de variables auxiliaires qui contient les variables du plan de sondage et  $x$ . Pour les plans de Poisson, nous avons utilisé  $z_i = (1, p_i, x_i, (1 - p_i)^{-1} p_i^i)'$  comme vecteur des variables d'équilibrage et comme vecteur des variables de régression. La première variable fournit un contrôle pour la taille de population, la deuxième est un contrôle pour la taille d'échantillon, la

parce que l'équilibrage est, par conception, semblable à la régression. Lorsque l'on utilise l'estimateur par la régression, le biais de ce dernier est du même ordre sous la méthode du cube que sous la méthode réjective. Pour l'échantillonnage réjectif, Fuller (2009a) donne les conditions de convergence de l'estimateur de variance de l'estimateur par la régression. Pour l'échantillonnage par la méthode du cube, Deville et Tillé (2005), ainsi que Tillé (2006) laissent entendre qu'en utilisant l'estimateur de variance pour un estimateur par la régression, on obtient une bonne approximation de la variance de l'estimateur de Horvitz-Thompson. Les estimateurs de variance proposés par Deville et Tillé (2005) donnent de bons résultats quand les probabilités d'inclusion conjointe du plan de sondage obtenu par la méthode du cube sont approximativement égales aux probabilités d'inclusion conjointe d'un plan de sondage de Poisson. Dans les études par simulation de la section 4, nous évaluons les estimateurs de variance proposés dans Fuller (2009a) et dans Deville et Tillé (2005).

#### 4. Simulation de l'estimateur par la régression

Une population de taille 100 a été générée à partir du modèle

$$y_i = x_i + 0,55x_i^2 + x_i e_i \quad (9)$$

où les  $x_i$  sont des valeurs fixes dans l'intervalle de 0 à 4 (figure 4). Soixante-douze des valeurs de  $x$  étaient des valeurs inférieures à 1,15 simulées aléatoirement d'après une loi exponentielle standard. Les 28 autres valeurs, variant de 0,18 à 4,0, ont été ajoutées de manière déterministe pour former l'ensemble de données de  $x$ . Les valeurs fixes de  $x$  ont été sélectionnées de manière que leur distribution soit assez étalée vers la droite afin que des grandes et des petites strates soient produites lors de la stratification de la population sur  $x$  de façon que les sommes intra-strate des valeurs  $x_i$  triées soient approximativement égales. La population a été maintenue fixe après la sélection initiale. Nous avons choisi le modèle (9), qui contient un terme quadratique, pour simuler la performance de la stratégie de sondage et d'estimation sous l'hypothèse du modèle (4) pour le plan de sondage et l'estimation.

Nous prenons pour plans de sondage initiaux l'échantillonnage de Poisson et l'échantillonnage aléatoire stratifié avec deux unités par strate. Nous avons déterminé les strates en fixant les limites de manière que la somme intra-strate des valeurs triées de  $x_i$  soit à peu près la même dans toutes les strates. Nous avons fixé la taille d'échantillon à 20 et formé dix strates. Les tailles de strate étaient 35, 15, 11, 9, 8, 7, 5, 4, 3 et 3. Pour la méthode réjective, nous avons utilisé un échantillonnage stratifié de deux unités par strate avec



de taille 20 tirés d'une population de 100 (figure 2). Pour l'échantillonnage aléatoire simple, la probabilité d'inclusion conjointe est 0,038. Sous échantillonnage réjectif, les probabilités d'inclusion conjointe sont approximativement égales à une fonction quadratique de  $x_i + x_j$ . Le tracé des probabilités d'inclusion conjointe sous échantillonnage par la méthode du cube en fonction de  $x_i + x_j$  semble présenter des angles plus aigus que celui des probabilités d'inclusion conjointe sous échantillonnage réjectif. Les probabilités d'inclusion conjointe élevées observées pour la méthode du cube sont associées à des paires d'unités situées sur les côtés opposés éloignés de  $\bar{x}_N$ . Autrement dit, pour la valeur d'échantillon de  $x_i + x_j$ , les paires dont la valeur  $|x_i| + |x_j|$  est grande ont une grande probabilité d'inclusion (figure 3).

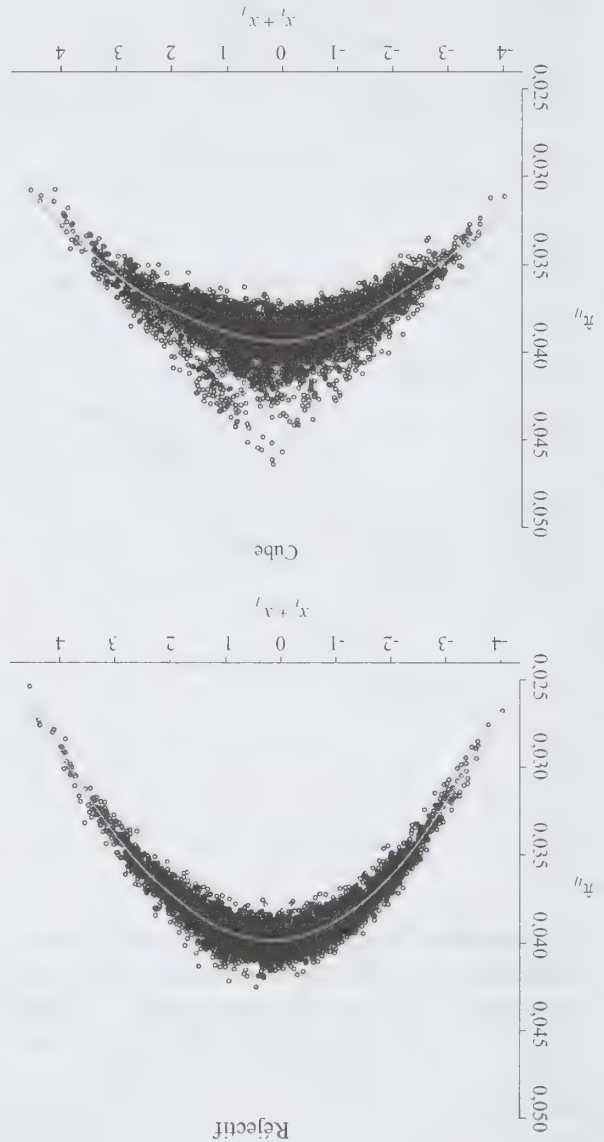
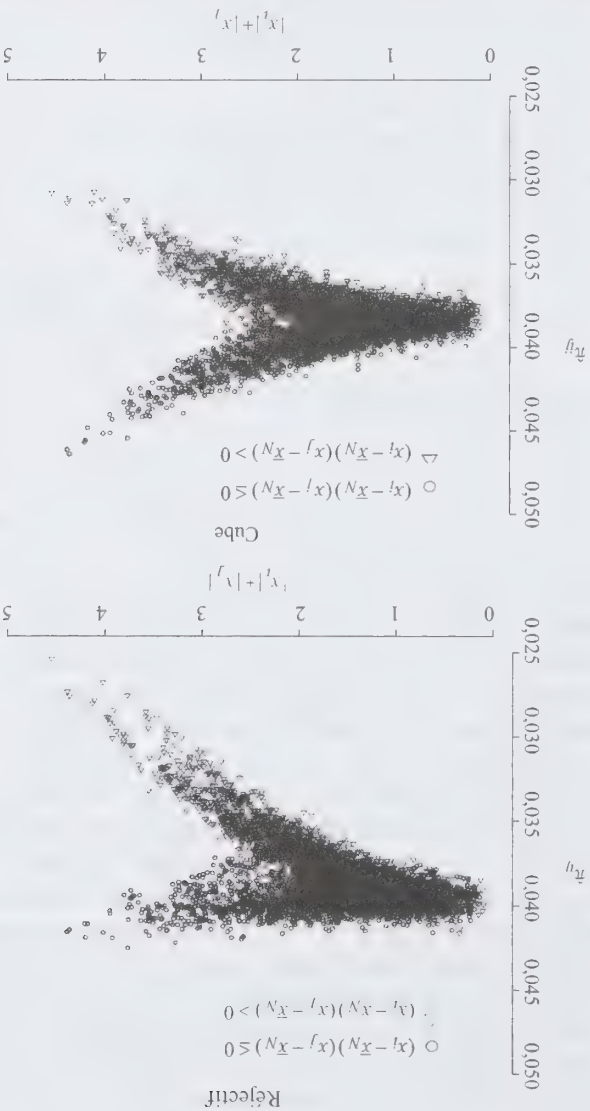


Figure 2 Probabilités d'inclusion de deuxième ordre simulées. Pour la méthode réjective, la variable d'échantillonnage est  $z_i = x_i$  et pour la méthode du cube,  $z_i = (p_i, x_i)'$ , où  $p_i = 20/100$

Figure 3 Probabilités d'inclusion de deuxième ordre simulées avec les sommes absolues de  $x$ . Pour la méthode réjective, la variable d'échantillonnage est  $z_i = x_i$  et pour la méthode du cube,  $z_i = (p_i, x_i)'$ , où  $p_i = 20/100$



L'estimateur de Horvitz-Thompson calculé en utilisant les probabilités d'inclusion initiales présente un biais d'ordre  $O_p(n^{-1})$  sous échantillonnage réjectif, tandis qu'il est sans biais sous échantillonnage par la méthode du cube. L'estimateur de variance de Horvitz-Thompson standard présente un biais sous les deux procédures. En utilisant des méthodes Monte Carlo, il est possible d'estimer des probabilités d'inclusion permettant d'utiliser des estimateurs de Horvitz-Thompson presque sans biais. Cependant, pour une grande population, il est difficile de simuler un nombre d'échantillons suffisant pour obtenir une estimation précise de la probabilité d'inclusion conjointe pour une paire d'unités. Au lieu d'estimer la variance, on peut utiliser un estimateur par la régression et l'estimateur de variance pour ce dernier. Cette approche est intuitivement séduisante,



d'échantillonnage obtenu par la méthode du cube. Comme prévu, la figure 1 indique que la méthode du cube maintient les probabilités d'inclusion de premier ordre spécifiées, mais que l'échantillonnage réjectif ne le fait pas. Par conséquent, l'estimateur de Horvitz-Thompson calculé en utilisant les probabilités d'inclusion initiales ( $p_i$ ) et les échantillons réjectifs est biaisé.

### 3. Probabilités d'inclusion

trissage dans les programmes.

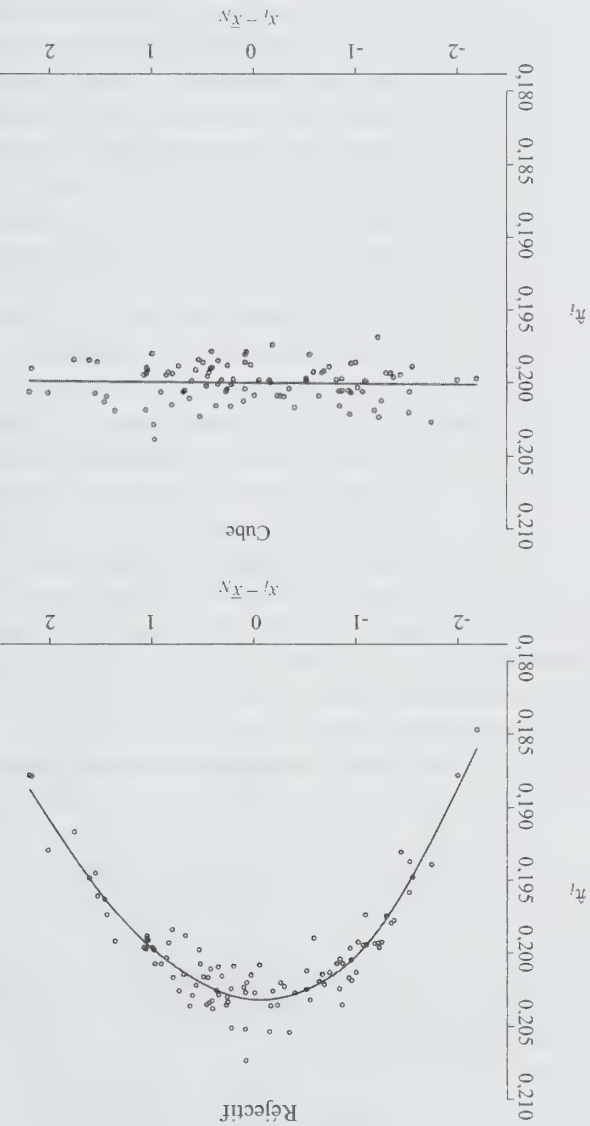
dans le présent article. Comme l'étape de minimisation du coût de l'échantillonnage par la méthode du cube requiert d'importantes ressources informatiques si l'on traite plus de 20 variables d'équilibrage, nous recommandons d'ajouter une étape de suppression de variables à la phase d'atter-

Soit  $\pi_i$  la probabilité d'inclusion de premier ordre de l'unité  $i$  et  $\pi_{ij}$  la probabilité d'inclusion conjointe des unités  $i$  et  $j$  sous un plan de sondage équilibré. Les probabilités d'inclusion de premier ordre initiales sont des données d'entrée requises tant pour l'échantillonnage réjectif que pour l'échantillonnage par la méthode du cube. Dans le cas de l'échantillonnage réjectif, les probabilités d'inclusion de premier ordre diffèrent des valeurs initiales, les unités proches de la moyenne de population ayant, sous cette méthode, une probabilité d'inclusion légèrement plus élevée que les unités éloignées de la moyenne. Par contre, dans l'échantillonnage par la méthode du cube, les probabilités d'inclusion de premier ordre demeurent celles de la spécification initiale. Autrement dit, pour l'échantillonnage par la méthode du cube,  $\pi_i = p_i$ . Bien que, pour l'échantillonnage réjectif,  $\pi_i \neq p_i$  en général, les estimateurs pris en considération utiliseront  $p_i$  plutôt que  $\pi_i$ .

Afin d'illustrer les différences entre les probabilités d'in-

clusion initiales et finales, nous avons simulé des échantillons de taille 20 tirés d'une population de 100 unités. La population de valeurs de  $x$  a été générée sous forme de variables aléatoires tirées d'une loi normale standard. La méthode de rejet s'appuyait sur l'échantillonnage aléatoire simple comme plan de sondage initial et était équilibrée sur  $x$ . Pour l'échantillonnage par la méthode du cube, nous avons utilisé un vecteur d'équilibrage de  $z_i = (p_i, x_i)$ , où  $p_i = 20/100$  pour tout  $i$ . L'inclusion de  $p_i$  dans le vecteur d'équilibrage pour l'échantillonnage par la méthode du cube avait pour but de contrôler la taille d'échantillon afin que le plan de sondage résultant soit comparable à l'utilisation de l'échantillonnage aléatoire simple comme plan initial dans la simulation de l'échantillonnage réjectif. Nous avons estimé les probabilités d'inclusion de premier ordre en utilisant une simulation Monte Carlo de taille 100 000 (figure 1). La courbe a été obtenue par un ajustement non paramétrique. Pour l'échantillonnage réjectif, nous avons utilisé un taux de rejet d'environ 90 %. En vertu de la théorie de l'échantillonnage réjectif, les probabilités d'inclusion de premier ordre correspondent approximativement à une fonction quadratique de la distance  $x_i - \bar{x}_N$  pour un plan de sondage initial à probabilités égales (Fuller 2009a). Le graphique donne à penser que toutes les probabilités d'inclusion de premier ordre sont égales à 0,2 pour le plan

**Figure 1** Probabilités d'inclusion de premier ordre simulées. Pour la méthode réjective, la variable d'équilibrage est  $z_i = x_i$  et pour la méthode du cube,  $z_i = (p_i, x_i)$ , où  $p_i = 20/100$



Dans la procédure d'échantillonnage réjectif, les probabilités d'inclusion conjointe diffèrent de celles du plan de sondage initial. Une paire d'unités  $i$  et  $j$  sont peu susceptibles d'avoir une probabilité d'inclusion conjointe élevée si  $x_i + x_j - 2\bar{x}_N$  est proche de zéro pour un plan de sondage initial à probabilités égales. Nous avons estimé les probabilités d'inclusion conjointe pour des échantillons simulés

## 2. Méthodes d'échantillonnage équilibré

L'échantillonnage réjectif comprend le rejet de tout échantillon qui ne satisfait pas une tolérance d'équilibrage spécifiée. Fuller (2009a) donne une condition pour le rejet d'un échantillon, tandis que Royall et Herson (1973) en présentent une autre. Dans la méthode de Fuller avec le vecteur de variables d'équilibrage  $\mathbf{z}_i$ , un échantillon tiré sous un plan de sondage initial est retenu si

$$(\bar{\mathbf{z}}_{\text{HT}} - \bar{\mathbf{z}}_N)' [V(\bar{\mathbf{z}}_{\text{HT}} | F_N)]^{-1} (\bar{\mathbf{z}}_{\text{HT}} - \bar{\mathbf{z}}_N) > \gamma \quad (8)$$

pour une constante donnée  $\gamma > 0$ , où  $\bar{\mathbf{z}}_{\text{HT}}$  est l'estimateur de Horvitz-Thompson de la moyenne pour la variable  $\mathbf{z}$ ,  $F_N$  est la population finie donnée,

$$V(\bar{\mathbf{z}}_{\text{HT}} | F_N) = N^{-2} \sum_N \sum_{j=1}^I (d_j - p_j d_j) \mathbf{z}_j \mathbf{z}_j' p_j^{-1} d_j^{-1},$$

$p_j$  est la probabilité d'inclusion de l'unité  $i$  et  $p_j$  est la probabilité d'inclusion conjointe de l'unité  $i$  et de l'unité  $j$  sous le plan initial. Autrement, l'échantillon est rejeté, un nouvel échantillon est tiré sous le plan de sondage initial et le respect de la condition (8) est vérifié pour le nouvel échantillon. Si le plan de sondage original obéit au théorème de la limite centrale, le premier membre de (8) est asymptotiquement une variable aléatoire  $\chi^2$  dont le nombre de degrés de liberté est égal au nombre de variables auxiliaires. Un taux de rejet approximatif peut être fixé en utilisant les quantiles d'une loi du  $\chi^2$  pour  $\gamma$ . Le choix du taux de rejet dépendra des objectifs particuliers de chaque enquête. Un faible taux de rejet pourrait ne pas réduire fortement la variance, mais donner à un chercheur un degré de certitude suffisant de ne pas sélectionner un échantillon très médiocre. Par ailleurs, un taux de rejet élevé pourrait réduire considérablement la variance, mais produire des tailles d'échantillon insuffisantes pour l'exécution d'analyses de domaine non planifiées. Par exemple, si un chercheur décide de procéder à une analyse de domaine sur la queue de la distribution d'une variable d'équilibrage, les probabilités d'inclusion conjointe pourraient être faibles, si bien que, pour de nombreux échantillons, le domaine ne contiendra que quelques unités.

La méthode du cube a été élaborée par Tillé et Deville, et décrite dans Tillé (2006). Elle a pour objectif de tirer un échantillon équilibré en se servant de probabilités d'inclusion de premier ordre prédéterminées. Si le vecteur des probabilités d'inclusion de premier ordre ne produit pas de plan équilibré, une étape supplémentaire consistant à minimiser une contrainte de coût est utilisée. Contrairement à la méthode réjective, les probabilités d'inclusion initiales d'ordre plus élevé ne sont pas spécifiées préalablement. L'étape de minimisation du coût assure le maintien des probabilités d'inclusion de premier ordre initiales spécifiées.

Pour faciliter la compréhension de la méthode du cube,

Tillé (2006) décrit l'échantillonnage géométriquement. L'en-semble de tous les échantillons possibles est défini comme étant l'ensemble des vecteurs correspondant aux sommets d'un hypercube unité  $N$ -dimensionnel, ou  $N$ -cube. Par exemple, si  $N=3$ , le sommet  $(0, 1, 1)$  désigne un échantillon contenant la deuxième et la troisième unité. Un plan d'équilibrage est créé en utilisant l'équation d'équilibrage (6) et la probabilité d'inclusion souhaitée  $p_i$  pour  $i = 1, \dots, N$ . Tout échantillon situé à l'intersection du plan d'équilibrage et d'un sommet de l'hypercube unité  $N$ -dimensionnel est équilibré. Le plan de sondage est équilibré si chaque point de l'intersection entre le plan d'équilibrage et l'hypercube unité est un sommet de ce cube. La procédure d'échantillonnage par la méthode du cube débute par la sélection d'un vecteur dans le plan d'équilibrage; cette étape est suivie par une marche aléatoire du point initial jusqu'à une arête de l'hypercube unité. Tillé donne à l'étape de la marche aléatoire le nom de phase de vol. Si le point rencontré sur l'arête à la fin de la marche aléatoire est un sommet de l'hypercube unité, l'échantillon est sélectionné. Sinon, une méthode de minimisation des coûts est utilisée pour convertir les composantes fractionnaires du vecteur d'arêtes en nombres entiers. Les composantes entières du vecteur d'arêtes ne sont pas modifiées durant l'étape de la minimisation du coût. Tillé donne à l'étape de minimisation du coût le nom de phase d'atterrissage. L'échantillonnage réjectif avec taux de rejet élevé produit des résultats semblables à l'échantillonnage par la méthode du cube.

D'autres méthodes que l'échantillonnage réjectif et l'échantillonnage par la méthode du cube peuvent être utilisées pour obtenir des échantillons presque équilibrés. Par exemple, la stratification où les limites des strates sont déterminées par les variables  $\mathbf{x}$  peut également avoir certains effets équilibrants sur les échantillons (Fuller 1981). Le processus suivi pour décider du nombre de variables qu'il convient d'utiliser dans les méthodes d'échantillonnage réjectif et d'échantillonnage par la méthode du cube est essentiellement le même que celui utilisé pour décider du nombre de variables à inclure dans un estimateur par la régression.

Un logiciel a été développé pour le tirage d'échantillons par la méthode du cube. Dans le cas de l'échantillonnage réjectif, des logiciels standard peuvent être utilisés pour tirer un échantillon et calculer (8). Une boucle doit être rédigée pour achever la procédure. Des programmes de tirage d'échantillons par la méthode du cube ont été écrits pour SAS et R. Voir Rousseau et Tardieu (2004) pour SAS, et Matei et Tillé (2005) pour R, ainsi que Deville et Tillé (2004), pour renseignements détaillés sur les procédures implémentées. Le programme R disponible dans la bibliothèque *sampling* a été utilisé pour les simulations décrites



meilleurs résultats que l'une ou l'autre technique utilisée seule.

Les plans de sondage équilibrés ont une certaine valeur pratique supplémentaire. Dans le cas de nombreux plans, il existe une probabilité non nulle de sélectionner un échantillon contenant des valeurs indésirables pour les variables

auxiliaires. Ainsi, un échantillon indésirable pourrait être un échantillon dont la répartition est insuffisante pour les domaines ou un échantillon présentant un grand nombre de valeurs extrêmes pour les variables auxiliaires. Même si les plans stratifiés réduisent l'ensemble d'échantillons éventuels de ce genre en fixant la taille d'échantillon dans chaque strate, l'obtention d'échantillons indésirables demeure possible. Par exemple, certains échantillons stratifiés pourraient être associés à certains poids négatifs à cause de l'utilisation d'estimateurs par la régression. L'équilibrage peut éliminer les échantillons donnant de mauvais résultats en retenant uniquement ceux qui produisent des estimations proches des quantités connues et ne possédant que des poids positifs pour les estimateurs par la régression.

L'échantillonnage équilibré a été proposé par Royall et Cumberland (1981) comme moyen de réduire le biais sous le modèle causé par la spécification incorrecte des modèles de superpopulation polynomiaux. Valliant, Dorfman et Royall (2000) discutent des incidences de l'équilibrage sous l'angle d'une approche prédictive de l'échantillonnage. Deville et Tillé (2004) ont étudié des méthodes de sélection d'échantillons équilibrés dans le cadre fondé sur le plan de sondage décrit plus haut. Le lecteur est invité à consulter également Tillé (2006, chapitre 8) pour un traitement détaillé de l'équilibrage. En pratique, il est parfois impossible de trouver un plan parfaitement équilibré. Un équilibrage très strict peut produire un plan présentant certaines probabilités d'inclusion conjointes extrêmes, y compris des probabilités d'inclusion nulles. Par conséquent, en pratique, on procède à un équilibrage partiel.

Dans le présent article, nous comparons, au moyen d'études par simulation, les propriétés des plans obtenus en appliquant deux méthodes d'équilibrage, à savoir l'échantillonnage réjectif de Fuller (2009a) et l'échantillonnage par la méthode du cube de Tillé (2006). Nous présentons aussi des modifications de la méthode d'échantillonnage réjectif de Fuller qui donnent plus de souplesse à l'équilibrage. À la section 2, nous décrivons l'échantillonnage réjectif et l'échantillonnage par la méthode du cube. À la section 3, nous comparons les propriétés des probabilités d'inclusion des deux méthodes d'équilibrage. À la section 4, nous présentons certains résultats de simulation obtenus en utilisant les échantillons équilibrés. À la section 5, nous donnons les corrections apportées à la méthode réjective. Enfin, à la section 6, nous présentons nos conclusions.

Si l'on dispose d'information auxiliaire au niveau de l'unité, on peut également l'intégrer dans le plan d'échantillonnage. Par exemple, dans un cas classique, on suppose que le modèle donné par

$$y_i = \beta_0 + \beta_1 x_i + x_i \varepsilon_i, \quad (4)$$

$\varepsilon_i \sim \text{ind}(0, \sigma^2)$  et  $\text{cov}(\varepsilon_i, x_j) = 0$  s'applique à la population  $F^N$ . Selon Isaki et Fuller (1982), les probabilités d'inclusion optimales pour l'estimateur par la régression sont celles qui sont proportionnelles à la racine carrée des variances sous le plan, c'est-à-dire  $P_i \propto x_i$  dans le cas qui nous occupe. Une méthode d'échantillonnage possible est l'échantillonnage de Poisson avec les probabilités d'inclusion

$$P_i = \left( \sum_{i=1}^N x_i \right)^{-1} x_i^n, \quad (5)$$

où  $n = \sum_{i=1}^N P_i$  est une taille d'échantillon cible spécifiée. Un deuxième plan d'échantillonnage fréquent, si l'on émet l'hypothèse du modèle (4), consiste à stratifier la population en se basant sur  $x$ . Les strates sont déterminées en fixant leurs limites de sorte que la somme des valeurs ordonnées de  $x_i$  créées dans chacune soit à peu près la même dans toutes les strates. Un nombre égal d'unités est sélectionné dans chaque strate. Ce plan de stratification, sous lequel les probabilités d'inclusion sont proches de (5), s'est avéré donner une variance anticipée proche de la meilleure variance sous le modèle d'échantillonnage par choix raisonné dans le cas de deux unités par strate (Fuller 1981).

Un autre moyen d'intégrer dans le plan de sondage l'information issue d'une variable auxiliaire est l'équilibrage. Un échantillon  $A$  est équilibré pour la variable  $z$  si

$$\bar{z}^{\text{HT}} = N^{-1} \sum_{i=1}^N P_i^{-1} z_i = N^{-1} \sum_{i=1}^N z_i = \bar{z}^N. \quad (6)$$

Un plan de sondage est équilibré pour  $z$  si chaque échantillon dont la probabilité est positive est équilibré pour  $z$ . L'équilibrage peut être considéré comme un calage par conception. Pour illustrer l'effet de l'équilibrage, considérons un plan avec probabilités d'inclusion égales et  $z_i = (1, x_i)'$ . La variance de prédiction conditionnelle de  $\bar{y}^{\text{reg}}$  sous le modèle (4) est donnée par

$$V(\bar{y}^{\text{reg}} - \bar{y}^N | \mathbf{x}, \bar{\mathbf{x}}^{\text{HT}}) = E\{V(\bar{u}^{\text{HT}} | F^N) | \mathbf{x}, \bar{\mathbf{x}}^{\text{HT}}\} + (\bar{\mathbf{x}}^N - \bar{\mathbf{x}}^{\text{HT}})' V(\beta_1 | \mathbf{x}, \bar{\mathbf{x}}^{\text{HT}}). \quad (7)$$

où  $u_i = x_i \varepsilon_i$ . Pour un plan équilibré, le deuxième terme de (7) est 0, ce qui laisse entendre que nous pourrions améliorer l'estimateur en équilibrant sur  $x$ . En pratique, une combinaison d'équilibrage et de calage donnera souvent de



# Comparaison de méthodes de restriction de l'ensemble d'échantillons

Jason C. Legg et Cindy L. Yu<sup>1</sup>

## Résumé

Dans le cas de nombreux plans de sondage, la probabilité de sélectionner un échantillon qui produira de mauvaises estimations pour des quantités connues n'est pas nulle. L'échantillonnage aléatoire stratifié permet de réduire l'ensemble de ces échantillons éventuels en fixant la taille de l'échantillon dans chaque strate. Cependant, l'obtention d'échantillons indésirables demeure possible après la stratification. L'échantillonnage réjectif permet d'éliminer les échantillons de mauvais résultats en ne retenant un échantillon que si des fonctions spécifiées des estimations sont comprises entre des limites de tolérance par rapport aux valeurs connues. Les échantillons résultant sont souvent dits équilibrés sur la fonction des variables utilisées dans la méthode de rejet. Nous présentons des modifications de la méthode de Fuller (2009a) qui donnent plus de souplesse aux règles de rejet. Au moyen de simulations, nous comparons les propriétés des estimations obtenues en suivant une méthode d'échantillonnage réjectif, d'une part, et une procédure d'échantillonnage par la méthode du cube, d'autre part.

Mots clés : Échantillonnage réjectif ; échantillonnage par la méthode du cube ; stratification ; échantillonnage équilibré.

## 1. Introduction

En échantillonnage, une pratique courante consiste à utiliser l'information de population connue au sujet des variables auxiliaires pour améliorer les estimateurs des moyennes et des totaux des caractéristiques d'intérêt. Lorsque l'on dispose de moyennes ou de totaux de population de contrôle pour une variable auxiliaire, on se sert souvent d'estimateurs par la régression et d'autres estimateurs par calage. Soit  $(x_i, y_i, d_i)$ ,  $i = 1, 2, \dots, N$ , une suite de vecteurs réels, où chaque  $x_i$  est un vecteur de dimension  $k$ , et un échantillon  $A$ , tiré de  $F_N = [(x_1, y_1, d_1), \dots, (x_N, y_N, d_N)]$  en utilisant un plan de sondage avec probabilités d'inclusion  $d_i$  et probabilités d'inclusion conjointe  $p_{ij}$ . Supposons que la moyenne de population de  $x_i$ ,  $\bar{x}_N$ , est connue. Considérons l'estimateur par la régression de la moyenne de population de la forme

$$\bar{y}^{\text{reg}} = \bar{z}_N' \hat{\beta}, \quad (1)$$

où  $z_i$  contient les variables du plan de sondage et  $x_i$ ,  $\bar{x}_N$  est la moyenne de population de  $z_i$  et  $\hat{\beta}$  est un estimateur des coefficients de régression. Pour de nombreux plans de sondage, l'estimateur  $\hat{\beta}$  de la forme

$$\hat{\beta} = \left( \sum_{i \in A} z_i \phi_i d_i^{-1} z_i' \right)^{-1} \sum_{i \in A} z_i \phi_i d_i^{-2} y_i, \quad (2)$$

où les valeurs de  $\phi_i$  sont des constantes déterminées par le plan, sera asymptotiquement efficace. Certains exemples de choix de  $\phi_i$  sont  $\phi_i = (1 - d_i)$  sous échantillonnage de Poisson et  $\phi_{hi} = (N_h - 1)^{-1} (N_h - n_h)$  pour l'élément  $i$  dans

$$\hat{\beta}_N = \left( \sum_{i=1}^N z_i \phi_i d_i^{-1} z_i' \right)^{-1} \sum_{i=1}^N z_i \phi_i d_i^{-1} y_i. \quad (2)$$

(2) converge avec pour tout  $i$ , l'estimateur (1) est convergent sous le plan (Fuller 2002). L'estimateur des coefficients de régression

$$\phi_i d_i^{-2} z_i' \hat{\beta} = d_i^{-1} \quad (3)$$

supposons qu'il existe un vecteur  $\beta$  tel que la strate  $h$  sous échantillonnage aléatoire stratifié. Si nous

Pour illustrer l'application de l'équation (3), supposons que nous prévoyons sélectionner un échantillon de Poisson et que nous voulons calculer la régression sur une seule covariable  $x_{1i}$  passant par l'origine. Si nous ajoutons  $(1 - d_i)^{-1} d_i$  dans  $z_i$  pour faire  $z_i' = (x_{1i}, [1 - d_i]^{-1} d_i)$ , l'estimateur (1) sera convergent sous le plan pour  $\bar{y}_N$  puisque l'expression (3) est satisfaite en prenant  $\beta' = (0, 1)$ . Si nous supposons en outre qu'une colonne de valeurs 1 se trouve dans l'espace colonnes des variables de régression  $z_i$ , alors pour ces valeurs de  $\phi_i$ , l'estimateur (1) atteint presque la variance asymptotique minimale pour des estimateurs par la régression convergents sous le plan sous certaines conditions de régularité (Rao 1994). Une alternative à la construction d'un estimateur par la régression consiste à partir d'un estimateur convergent sous le plan, tel que l'estimateur par la régression généralisée de Särndal (1980) et à déterminer le meilleur coefficient sachant cette forme de l'estimateur. En débutant avec une forme convergente sous le plan, il n'est plus nécessaire de satisfaire l'expression (3). La condition (3) permet d'exprimer l'estimateur (1) sous la

1. Cindy L. Yu est professeure adjointe au Département de statistique et au Center for Survey Statistics and Methodology à la Iowa State University, Ames, IA 50010. Courriel : cindy.yu@iastate.edu ; Jason C. Legg est un chercheur postdoctoral au Center for Survey Statistics and Methodology à la Iowa State University, Ames, IA 50010. Courriel : jason-legg@hotmail.com.

Si nous supposons que  $\hat{\beta}_1(j) \approx \beta_1$ ,  $\hat{\beta}_2(j) \approx \beta_2$ , et comme Rao et Sitter (1995), que  $\bar{x}_n(j)/\bar{x}_p(j) \approx \bar{x}_n/\bar{x}_p$ , il est assez simple de montrer que

$$\begin{aligned} \sum_{j \in \lambda_1} [\hat{Y}_1^*(j) - \hat{Y}_1]^2 &\approx \sum_{j \in \lambda_1} [\hat{Y}_1^*(j) - \hat{Y}_1^*]^2 + \hat{\beta}_2^2 \sum_{j \in \lambda_1} [\hat{X}_1^*(j) - \hat{X}_1^*]^2 \\ &+ 2\hat{\beta}_2 \sum_{j \in \lambda_1} [\hat{Y}_1^*(j) - \hat{Y}_1^*][\hat{X}_1^*(j) - \hat{X}_1^*] \\ &- 2\hat{\beta}_2 \sum_{j \in \lambda_1} [\hat{Y}_1^*(j) - \hat{Y}_1^*][\hat{X}_2^*(j) - \hat{X}_2^*] \\ &- 2\hat{\beta}_2^2 \sum_{j \in \lambda_1} [\hat{X}_1^*(j) - \hat{X}_1^*][\hat{X}_2^*(j) - \hat{X}_2^*] \\ &- 2\hat{\beta}_1 \hat{\beta}_2 \sum_{j \in \lambda_1} [\hat{Y}_1^*(j) - \hat{Y}_1^*][\hat{Z}_1^*(j) - \hat{Z}_1^*] \\ &- 2\hat{\beta}_1 \hat{\beta}_2^2 \sum_{j \in \lambda_1} [\hat{X}_1^*(j) - \hat{X}_1^*][\hat{Z}_1^*(j) - \hat{Z}_1^*] \\ &+ \hat{\beta}_2^2 \sum_{j \in \lambda_1} [\hat{X}_1^*(j) - \hat{X}_1^*]^2 \\ &+ \hat{\beta}_1^2 \sum_{j \in \lambda_1} [\hat{Z}_1^*(j) - \hat{Z}_1^*]^2 \\ &- 2\hat{\beta}_1 \hat{\beta}_2^2 \sum_{j \in \lambda_1} [\hat{X}_1^*(j) - \hat{X}_1^*][\hat{Z}_1^*(j) - \hat{Z}_1^*] \\ &- 2\hat{\beta}_1 \hat{\beta}_2 \sum_{j \in \lambda_1} [\hat{X}_1^*(j) - \hat{X}_1^*][\hat{Z}_1^*(j) - \hat{Z}_1^*] \end{aligned}$$

Puisque les dix termes du deuxième membre de cette équation pour  $\sum_{j \in \lambda_1} [\hat{Y}_1^*(j) - \hat{Y}_1^*]^2$  sont les estimateurs convergents des dix termes analogues de l'équation susmentionnée pour  $V(\hat{Y}_1^*)$ , nous pouvons conclure que l'estimateur jackknife de variance (2.2) est convergent.

## Bibliographie

- Ahmed, M.S. (1997). The general class of chain estimators for the ratio of two means using double sampling. *Communications in Statistics. Theory and Methods*, 26(9), 2247-2254.
- Arnab, R., et Singh, S. (2006). A new method for estimating variance from data imputed with ratio method of imputation. *Statistics and Probability Letters*, 76, 513-519.
- Berger, Y. (2007). A jackknife variance estimator for unstage stratified samples with unequal probabilities. *Biometrika*, 94, 953-964.
- Berger, Y., et Skinner, C. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society, Series B*, 67, 79-89.
- Chand, L. (1975). *Some ratio type estimators based on two or more auxiliary variables*. Thèse de doctorat, Iowa State University, Ames, Iowa, E.-U.
- Chen, J., et Shao, J. (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96, 260-269.
- Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 117-132.
- Hidiroglou, M.A., et Särndal, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Vol. II, 873-878.
- Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 11-20.
- Kim, J.K., Navarro, A., et Fuller, W.A. (2000). Variance estimation for 2000 Census coverage estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 515-520.
- Kim, J.K., Navarro, A., et Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.
- Kim, J.K., et Sitter, R.R. (2003). Efficient replication variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.
- Kott, P.S., et Shukel, D. (1997). La méthode du jackknife convient-elle à un échantillon à deux degrés ? *Techniques d'enquête*, 23, 89-98.
- Kovar, J., et Chen, E. (1994). Méthode du jackknife pour l'estimation de la variance en présence de données imputées. *Techniques d'enquête*, 20, 47-55.
- Raj, D. (1965). On sampling over two occasions with probability proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.
- Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-60.
- Singh, S. (2000). Estimation of variance of regression estimator in two phase sampling. *Calcutta Statistical Association Bulletin*, 50, 49-63.
- Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.
- Srivenkataramana, T., et Tracy, D.S. (1989). Two-phase sampling for selection with probability proportional to size in sample surveys. *Biometrika*, 76, 818-821.
- Statistique Canada, N° 12-001-X au catalogue





et

$$s_{d(t|k)}^2 = (n-1) \sum_{i=1}^n [Y_{it(t|k)} - \bar{Y}_{(t|k)} - r_{1(t|k)}(z_{it(t|k)} - \bar{z}_{(t|k)})]^2$$

avec  $r_{1(t|k)} = \bar{Y}_{(t|k)} / \bar{Z}_{(t|k)}$ . Nous calculons aussi l'estimateur jackknife de variance

$$\hat{V}_{\text{JACK}}[(\hat{Y}_{\text{ch}}^c(t|k))] = \frac{m-1}{m} \sum_{j=1}^m \left[ \frac{\bar{X}'_{(t|k)}(j)}{\bar{Z}_{(t|k)}(j)} \frac{\bar{X}_{(t|k)}(j)}{\bar{Z}_{(t|k)}(j)} - \bar{Y}_{(t|k)} \frac{\bar{X}'_{(t|k)}}{\bar{Z}_{(t|k)}} \right] \quad (4.10)$$

En nous servant de la même population générée de taille  $N = 500$ , nous répétons la simulation ; toutefois, cette fois-ci, nous utilisons  $m = 400$  et  $n = 80$ . Puis, nous créons trois populations supplémentaires de taille  $N = 500$  en utilisant  $\beta_1 = 0,5$  avec  $\beta_2 = 0,5$ ,  $\beta_1 = 3,5$  avec  $\beta_2 = 0,5$ , et  $\beta_1 = 0,5$  avec  $\beta_2 = 2,5$ . Pour chacune de ces trois populations, nous répétons les deux simulations décrites plus haut, où, dans la première simulation,  $m = 100$  avec

**Tableau 2**  
Comparaison des estimateurs jackknife et usuel de la variance de l'estimateur par le ratio en chaîne de la moyenne de population où la variable auxiliaire  $X$  suit une loi gamma dont le paramètre de forme est 2,2 et le paramètre d'échelle, 1, et la variable auxiliaire  $Z$  suit une loi gamma dont le paramètre de forme est 0,1 et le paramètre d'échelle, 1

$m$	$n$	$\beta_1$	$\beta_2$	$N$	RV	EBU	EBJ	ER
100	20	3,5	2,5	500	0,769	0,000	0,027	1,063
				5 000	0,831	-0,012	0,020	2,282
				50 000	0,818	-0,006	0,028	1,785
				500 000	0,852	0,001	0,036	1,993
				5 000	0,911	-0,001	0,004	0,791
				50 000	0,943	-0,001	0,002	0,888
				500 000	0,948	0,000	0,003	0,896
				5 000	0,946	0,000	0,003	0,899
				50 000	0,845	-0,001	0,015	1,674
				500 000	0,932	-0,011	0,004	3,632
				5 000	0,947	-0,005	0,004	3,221
				50 000	0,947	-0,005	0,004	3,221
				500 000	0,866	-0,001	0,009	0,668
				5 000	0,858	-0,003	0,008	0,775
				50 000	0,855	-0,001	0,010	0,670
				500 000	0,855	-0,001	0,012	0,697
				5 000	0,540	0,000	0,013	0,044
				50 000	0,780	-0,001	0,009	1,346
				500 000	0,819	0,000	0,008	1,878
				5 000	0,810	-0,001	0,006	1,953
				50 000	0,817	0,000	0,003	0,254
				500 000	0,956	0,000	0,000	0,885
				5 000	0,973	0,000	0,001	0,946
				50 000	0,973	0,000	0,000	0,963
				500 000	0,579	0,000	0,010	0,041
				5 000	0,907	-0,001	0,003	3,158
				50 000	0,954	0,000	0,002	3,845
				500 000	0,950	-0,001	0,001	4,853
				5 000	0,787	0,000	0,004	0,222
				50 000	0,862	0,000	0,002	0,570
				500 000	0,873	0,000	0,003	0,698
				5 000	0,875	0,000	0,002	0,595

Nous considérons aussi l'estimateur par la régression de Sitter (1997) et répétons entièrement l'étude par simulations exécutée en utilisant l'estimateur par le ratio (4.1). En particulier, au lieu de (4.1), nous utilisons l'estimateur

$$\hat{Y}_i^S = \bar{y} + b^*(\bar{x}' - \bar{x}) \quad (4.4)$$

dont la variance approximative est

$$V(\hat{Y}_i^S) = (n^{-1} - m^{-1})S_d^2 + (m^{-1} - N^{-1})S_y^2, \quad (4.5)$$

où

$$S_d^2 = (N - 1)^{-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - \beta^{\text{Pop}}(X_i - \bar{X})]^2$$

avec

$$\beta^{\text{Pop}} = \sum_{i=1}^N X_i Y_i / \sum_{i=1}^N X_i^2.$$

Pour chaque combinaison distincte de  $N$ ,  $g$ ,  $m$  et  $n$  utilisée dans l'étude par simulations, nous calculons

$$V[(\hat{Y}_i^S | t | k)] = (n^{-1} - m^{-1})S_d^{d(t|k)} + (m^{-1} - N^{-1})S_2^{v(t|k)}, \quad (4.6)$$

pour le  $t^{\text{e}}$  échantillon de deuxième phase provenant du  $k^{\text{e}}$  échantillon de première phase, où la variance d'échantillon est

$$S_2^{d(t|k)} = (n - 1)^{-1} \sum_{i=1}^n [(Y_{i(t|k)} - \bar{y}_{(t|k)}) - b_{*}^{(t|k)}(x_{i(t|k)} - \bar{x}_{(t|k)})]^2.$$

Nous calculons aussi l'estimateur jackknife de variance

$$V_{\text{JACK}}[(\hat{Y}_i^S | t | k)] = \frac{m}{m-1} \sum_{j=1}^m [\bar{y}_{(t|k)}^{(j)} + b_{*}^{(t|k)}(j) \{ \bar{x}_{(t|k)}'(j) - \bar{x}_{(t|k)}(j) \}]^2$$

$$- \{ \bar{y} + b^*(\bar{x}' - \bar{x}) \}^2. \quad (4.7)$$

Pour chaque combinaison distincte de  $N$ ,  $g$ ,  $m$  et  $n$ , nous utilisons les équations (4.5) à (4.7) pour calculer les valeurs de  $R_V$ ,  $EBU$ ,  $EBJ$  et  $ER$  correspondant à celles présentées dans le tableau 1 pour l'estimateur (4.1). Les résultats sont extrêmement semblables à ceux obtenus pour l'estimateur par le ratio.

## 4.2 Étude par simulations : estimateurs de Chand (1975) et d'Ahmed (1997)

Pour le deuxième ensemble de simulations, nous supposons que, quand l'échantillon de première phase de  $m$  unités est sélectionné dans la population de taille  $N$ , l'information sur deux variables auxiliaires  $X$  et  $Z$  est recueillie. Quand l'échantillon de deuxième phase de taille  $n$  est sélectionné dans l'échantillon de première phase, la variable

$$Y_i = \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$

unités d'observation  $(X_i, Z_i, Y_i)$  en utilisant Nous commençons par créer une population de  $N = 500$

étudiée  $Y$  est mesurée, ainsi que les deux variables auxiliaires  $X$  et  $Z$ . Nous supposons aussi que la variable auxiliaire  $Z$  est connue pour l'ensemble de la population.

avec  $\beta_1 = 3,5$  et  $\beta_2 = 2,5$ . Pour chaque  $i$ ,  $i = 1, \dots, N$ , nous tirons  $X_i$  d'une loi gamma dont le paramètre de forme est égal à 2,2 et le paramètre d'échelle, à 1,  $Z_i$  d'une loi gamma dont le paramètre de forme est égal à 0,1 et le paramètre d'échelle, à 1, et  $\varepsilon_i$  d'une loi normale standard. Parmi la population résultante d'observations  $(X_i, Z_i, Y_i)$ , nous sélectionnons 1 000 échantillons de première phase de  $m = 100$  unités et, dans chacun de ces échantillons, nous sélectionnons 10 000 échantillons de deuxième phase de  $n = 20$  unités.

Suivant Chand (1975), un estimateur par le ratio en chaîne sous échantillonnage à deux phases est donné par

$$\hat{Y}_i^{\text{Ch}} = \bar{y}(\bar{x}' / \bar{x})(\bar{Z} / \bar{z}),$$

dont la variance approximative est

$$V(\hat{Y}_i^{\text{Ch}}) = (n^{-1} - m^{-1})S_d^2 + (m^{-1} - N^{-1})S_2^{d_i}, \quad (4.8)$$

où

$$S_2^{d_i} = (N - 1)^{-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - R_2(X_i - \bar{X})]^2$$

et

$$S_2^{d_i} = (N - 1)^{-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - R_1(Z_i - \bar{Z})]^2$$

avec

$$\bar{Y} = \sum_{i=1}^N Y_i / N, \quad \bar{X} = \sum_{i=1}^N X_i / N, \quad \bar{Z} = \sum_{i=1}^N Z_i / N,$$

$R_1 = \bar{Y} / \bar{Z}$  et  $R_2 = \bar{Y} / \bar{X}$ . Dans l'étude par simulations, nous calculons

$$V[(\hat{Y}_i^{\text{Ch}} | t | k)] = (n^{-1} - m^{-1})S_2^{d_2(t|k)} + (m^{-1} - N^{-1})S_2^{d_1(t|k)} \quad (4.9)$$

pour le  $t^{\text{e}}$  échantillon de deuxième phase tiré du  $k^{\text{e}}$  échantillon de première phase, où les variances d'échantillon sont

$$S_2^{d_2(t|k)} = (n - 1)^{-1} \sum_{i=1}^n [(Y_{i(t|k)} - \bar{y}_{(t|k)}) - r_2(t|k)(x_{i(t|k)} - \bar{x}_{(t|k)})]^2$$

avec

$$r_2(t|k) = \bar{y}_{(t|k)} / \bar{x}_{(t|k)}$$

plus haut où, dans la première simulation,  $m = 100$  et  $n = 20$  et, dans la deuxième simulation,  $m = 400$  et  $n = 80$ . Enfin, pour étudier l'effet de la taille de la population, nous répétons toutes les simulations basées sur les trois valeurs de  $g$ ,  $m$  et  $n$  quand  $N = 500$  pour trois valeurs supplémentaires de  $N$ , à savoir 5 000, 50 000 et 500 000. Les résultats obtenus pour RV, EBU, EBJ et ER pour chacune de ces simulations sont présentés au tableau 1.

Dans ce tableau, les résultats pour l'efficacité relative (ER) donnent à penser que, quand la taille de population  $N$  tend vers l'infini (comme l'ont envisagé Rao et Sitter 1995), l'estimateur jackknife de variance demeure plus efficace que l'estimateur sans biais usuel de variance. Nous constatons aussi que, pour de très grandes valeurs de  $N$ , les valeurs de RV tendent vers 1. Cependant, si nous prenons les cas où  $N = 500$ , quand la taille de population est relativement faible, non seulement les valeurs de RV sont appréciablement

Tableau 1  
Comparaison des estimateurs jackknife et usuel de la variance de l'estimateur par le ratio de la moyenne de population quand  $\beta = 10$  et que la variable auxiliaire  $X$  suit une loi gamma avec un paramètre de forme de 3,1 et un paramètre d'échelle de 1

$N$	$m$	$n$	$g$	RV	EBU	EBJ	ER
500	100	20	0,0	0,801	0,006	0,542	1,521
			0,5	0,800	0,010	0,579	1,310
			1,0	0,805	-0,071	0,561	1,267
			1,5	0,816	-0,358	0,575	1,149
5 000	100	20	0,0	0,840	-0,720	1,777	0,935
			0,5	0,979	-0,028	0,042	4,015
			1,0	0,976	0,007	0,096	3,709
			1,5	0,965	0,023	0,172	3,210
50 000	100	20	0,0	0,916	-1,103	0,493	0,967
			0,5	1,001	-0,002	0,003	6,241
			1,0	0,998	0,107	0,126	4,936
			1,5	0,937	-0,211	0,167	1,558
500 000	100	20	0,0	0,924	-0,355	0,940	1,005
			0,5	1,001	-0,057	-0,054	4,730
			1,0	0,999	0,014	0,024	4,669
			1,5	0,993	0,185	0,229	3,223
5 000	400	80	0,0	0,907	-1,054	0,530	1,009
			0,5	0,214	0,000	0,520	0,002
			1,0	0,237	-0,001	0,523	0,002
			1,5	0,320	0,000	0,544	0,006
50 000	400	80	0,0	0,530	-0,001	0,616	0,066
			0,5	0,733	-0,012	1,091	0,452
			1,0	0,919	-0,003	0,061	2,687
			1,5	0,920	-0,001	0,064	2,505
500 000	400	80	0,0	0,930	-0,028	0,077	2,058
			0,5	0,940	-0,089	0,184	1,088
			1,0	0,991	0,000	0,012	5,276
			1,5	0,980	-0,024	-0,001	1,777
500 000	400	80	0,0	0,967	-0,171	-0,040	1,099
			0,5	1,000	0,009	0,009	5,501
			1,0	0,999	0,001	0,001	5,180
			1,5	0,993	-0,001	0,006	3,852
500 000	400	80	0,0	0,992	-0,022	-0,018	1,809
			0,5	0,971	-0,179	-0,079	1,136



avec  $\bar{Y} = \sum_{i=1}^N Y_i / N$ ,  $\bar{X} = \sum_{i=1}^N X_i / N$  et  $R = \bar{Y} / \bar{X}$ . Pour le  $t^e$  échantillon de deuxième phase ( $t = 1, \dots, 10\,000$ ), nous calculons l'estimateur usuel de variance

$$V[(\hat{Y}_{RS}^{(t)} | k)] = \left( \frac{1}{1} - \frac{1}{n} \right) s_{d(t|k)}^2 + \left( \frac{1}{1} - \frac{1}{N} \right) s_{y(t|k)}^2 \quad (4.2)$$

où les variances d'échantillon sont

$$s_{d(t|k)}^2 = (n-1)^{-1} \sum_{i=1}^n [(Y_{it(t|k)} - \bar{Y}_{(t|k)}) - R(Y_{it(t|k)} - \bar{X}_{(t|k)})]^2$$

et

$$s_{y(t|k)}^2 = (n-1)^{-1} \sum_{i=1}^n (Y_{it(t|k)} - \bar{Y}_{(t|k)})^2$$

avec  $\bar{Y}_{(t|k)} = \sum_{i=1}^n Y_{it(t|k)} / n$  et  $\bar{X}_{(t|k)} = \sum_{i=1}^n X_{it(t|k)} / n$ . En outre,  $r_{(t|k)} = \bar{Y}_{(t|k)} / \bar{X}_{(t|k)}$ . Nous calculons aussi l'estimateur jackknife de variance

$$V_{JACK}[(\hat{Y}_{RS}^{(t)} | k)] = \left[ \frac{m}{m-1} \sum_{j=1}^m \left( \frac{\bar{X}'_{(t|k)}(j)}{\bar{Y}'_{(t|k)}(j)} - \frac{\bar{X}_{(t|k)}}{\bar{Y}_{(t|k)}} \right) \right]^2 \quad (4.3)$$

et le ratio des variances estimées

$$RV(t|k) = V[(\hat{Y}_{RS}^{(t)} | k)] / V_{JACK}[(\hat{Y}_{RS}^{(t)} | k)].$$

Puis, nous calculons la moyenne de  $RV(t|k)$  sur tous les  $k$  et  $t$ , laquelle est donnée par

$$RV = \frac{1}{10\,000} \sum_{k=1}^{10\,000} \sum_{t=1}^{10\,000} RV(t|k).$$

Nous déterminons aussi les estimations empiriques des biais dans (4.2) et (4.3) en calculant

$$EBU = \frac{1}{10\,000} \sum_{k=1}^{10\,000} \sum_{t=1}^{10\,000} \{V[(\hat{Y}_{RS}^{(t)} | k)] - V(\hat{Y}_{RS}^{(t)})\},$$

et

$$EBJ = \frac{1}{10\,000} \sum_{k=1}^{10\,000} \sum_{t=1}^{10\,000} \{V_{JACK}[(\hat{Y}_{RS}^{(t)} | k)] - V(\hat{Y}_{RS}^{(t)})\}.$$

Notons que l'estimateur donné par (4.2) est sans biais. Enfin, nous calculons l'efficacité relative de l'estimateur usuel de variance par rapport à l'estimateur jackknife par

$$ER = \frac{\frac{1}{10\,000} \sum_{k=1}^{10\,000} \sum_{t=1}^{10\,000} \{V[(\hat{Y}_{RS}^{(t)} | k)] - V(\hat{Y}_{RS}^{(t)})\}^2}{\frac{1}{10\,000} \sum_{k=1}^{10\,000} \sum_{t=1}^{10\,000} \{V_{JACK}[(\hat{Y}_{RS}^{(t)} | k)] - V(\hat{Y}_{RS}^{(t)})\}^2}.$$

En nous servant de la même population générée de taille  $N = 500$ , nous répétons la simulation, mais en utilisant cette fois  $m = 400$  et  $n = 80$ . Puis, nous créons quatre populations supplémentaires de taille  $N = 500$  en utilisant  $g = 0,5, 1,0, 1,5$  et  $2,0$ . Pour chacune de ces quatre populations, nous répétons les deux simulations décrites

chaîne d'Ahmed (1997). Pour commencer, nous décrivons et présentons les résultats des simulations qui ont été effectuées pour les estimateurs de Sitter et Rao (1995) et de Sitter (1997). Suivent une discussion et un résumé de simulations semblables pour les estimateurs de Chand (1975) et d'Ahmed (1997). Puisque, contrairement au cas des estimateurs par le ratio et par la régression, de l'information complète sur une deuxième variable auxiliaire  $Z$  est requise pour l'ensemble de la population afin d'appliquer les deux estimateurs en chaîne, les simulations qui ont été effectuées pour ces estimateurs sont un peu plus compliquées que celles exécutées pour les estimateurs par le ratio et par la régression.

#### 4.1 Étude par simulations : estimateurs de Rao et Sitter (1995) et de Sitter (1997)

Pour le premier ensemble de simulations, nous supposons que l'échantillon de première phase de  $m$  unités est sélectionné dans une population de  $N$  unités et que seule la variable auxiliaire  $X$  est mesurée. Dans l'échantillon de première phase de  $m$  unités, nous sélectionnons ensuite un échantillon de deuxième phase de  $n$  unités par EASSR dans lequel sont mesurées la variable étudiée,  $Y$ , ainsi que la variable auxiliaire,  $X$ .

Nous commençons par créer une population de  $N$  unités constituée de paires  $(X_i, Y_i)$  en utilisant le modèle

$$Y_i = \beta X_i + \sqrt{X_i} \varepsilon_i,$$

avec  $\beta = 10$ . Au départ, nous fixons que  $g = 0$  et  $N = 500$ . Pour chaque  $i$ ,  $i = 1, \dots, N$ , nous générons  $X_i$  à partir d'une loi gamma dont le paramètre de forme est égal à 3,1 et le paramètre d'échelle est égal à 1, et  $\varepsilon_i$  suit une loi normale standard. Parmi la population résultante de paires  $(X_i, Y_i)$ , nous sélectionnons 1 000 échantillons de première phase de  $m = 100$  unités, et dans chacun de ces échantillons, nous sélectionnons 10 000 échantillons de deuxième phase de  $n = 20$  unités.

Sous le plan d'échantillonnage utilisé ici, Rao et Sitter (1995) ont proposé l'estimateur par le ratio

$$\hat{Y}_{RS}^{(t)} = \bar{y}(\bar{x}' / \bar{x}), \quad (4.1)$$

dont la variance approximative est

$$S_{y'}^2 = (N-1)^{-1} \sum_{i=1}^N [(Y_i' - \bar{Y}) - R(X_i' - \bar{X})]^2$$

et

$$S_y^2 = (N-1)^{-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$$

également sélectionné par EASSR. Manifestement,  $d_{i1} = N/m$  et  $d_{z1} = m/n$ , de sorte que  $w_{o11} = 1/m$  et  $w_{z1} = 1/n$ . Si  $q_{11} = 1/z_1$  et  $q_{z1} = 1/x_1$ , l'estimateur calé  $Y_c$  devient

devient

$$(3.10) \quad \underline{Y}_c^{\text{ch}} = \underline{y}(\underline{x}' / \underline{x})(\underline{z}' / \underline{z}),$$

no

$$u / {}^1x \sum_{! \in s_1} = , \underline{x} , u / {}^1x \sum_{! \in s_2} = \underline{x} , u / {}^1x \sum_{! \in s_2} = \underline{y}$$

et  $\bar{z}' = \sum_{i \in s_1} z_i / m$ . L'estimateur jackknife de  $Y$  est

$$(3.11) \quad \left. \begin{aligned} & \text{si } f \in s_{-1} \quad \frac{(f), \underline{z}}{\underline{z}} \frac{\underline{x}}{(f), \underline{x}} (f), \underline{y} \\ & \text{si } f \in s_z \quad \frac{(f), \underline{z}}{\underline{z}} \frac{(f), \underline{x}}{(f), \underline{x}'} (f), \underline{y} \end{aligned} \right\} = (f)_{\text{ch}}^{\text{ch}} I$$

on a  $\underline{p}(f) = (n\overline{y} - x_f)/(n - 1)$ ,  $\underline{x}(f) = (n\overline{x} - x_f)/(n - 1)$  et enfin  $\underline{z}(f) = (m\overline{z} - x_f)/(m - 1)$ . Si nous posons que  $R_1 = \underline{x}/\underline{z}'$  (un estimateur de  $R_1 = \underline{X}/\underline{Z}$ ) et que  $R_2 = \underline{p}/\underline{x}$  (un estimateur de  $R_2 = \underline{Y}/\underline{X}$ ), et similairement que  $R_1(f) = \underline{x}(f)/\underline{z}'(f)$  et  $R_2(f) = \underline{p}(f)/\underline{x}(f)$ , la différence entre (3.11) et (3.10) peut s'écrire

$$Y_c^{\text{ch}}(j) = \left\{ \begin{array}{l} \varepsilon_2(j) + \hat{R}_2 \varepsilon_1(j) + \hat{R}_2(d_2(j) + \hat{R}_2 \delta_2(j)) \quad \text{si } j \in s_2 \\ \hat{R}_2 \varepsilon_1(j) \quad \text{si } j \in (s_1 - s_2) \end{array} \right. \quad (3.12)$$

ou nous pouvons écrire dans (3.12) que  $e_2(j) = \bar{u}(j) - \bar{x}_2(j) - R_2(j)\{\bar{x}(j) - \bar{x}\} - R_1(j)R_2(j)\{\bar{z}(j) - \bar{Z}\}, d_2(j) = \bar{x}(j) - \bar{x}_1, \delta_2(j) = \{\bar{x}(j) - \bar{x}(j)\} - \{R_1(j)\{\bar{Z} - \bar{z}'(j)\} - R_1(j)\{\bar{Z} - \bar{z}'(j)\} - \bar{x}(j)\}$ , et, enfin, que le terme  $e_1(j) = \{\bar{x}(j) - \bar{x}\} - R_1(j)\{\bar{Z} - \bar{z}'(j)\}$ ,  $R_1(j)\{\bar{Z} - \bar{z}'(j)\}$ ,  $R_1(j)\{\bar{z}(j) - \bar{z}'(j)\}$ . Donc, l'estimateur jackknife de la variance de l'estimateur  $\hat{Y}_{ch}^{ch}$  est donné par

$$D_{\text{JACK}}(\hat{X}_{\text{c}}^{\text{Ch}}) = \left\{ m(1 - \mathbb{I})/m \left[ \sum_{j \in \mathcal{S}_2} \mathcal{E}_2^z(j) + \sum_{j \in \mathcal{S}_2} \hat{\mathcal{R}}_2^z(j) d_2^z(j) \right. \right. \\ \left. \left. + \hat{\mathcal{R}}_2^z(j) \delta_2(j) \{ \delta_2(j) + 2\epsilon_1(j) \} \right. \right. \\ \left. \left. + 2\hat{\mathcal{R}}_2(j) \mathcal{E}_1(j) \epsilon_2(j) \right. \right. \\ \left. \left. + 2\hat{\mathcal{R}}_2(j) d_2^z(j) \{ \mathcal{E}_1(j) + \delta_2(j) \} \right. \right. \\ \left. \left. + \sum_{j \in \mathcal{S}_1} \hat{\mathcal{R}}_2^z(j) \mathcal{E}_2^z(j) \right] \right\}.$$

Cas 3.6 : Ahmed (1997)

Considérons le même plan d'échantillonnage que pour le cas 3.5. Puisqu'il est évident que  $q_{11} = 1/z_1$  et  $q_{21} = 1/x_1$ , comme dans Chand (1975), nous posons que  $q_{11} = q_{21} = 1$  et  $q_{21} = 1/x_1$ , de sorte que l'estimateur calé se réduit à

$$\hat{Y}_i^{\text{Chit}} = \hat{y} + b_2^*(x' - x) + b_1^*z_2'(\bar{Z} - z'), \quad (3.13)$$

$$(3.14) \quad Y_c^{\text{Chp}}(f) = \left\{ \begin{array}{ll} \left\{ \underline{y}(f) + b_*^{\bar{z}_2}(f) \{ \underline{x}'(f) - \underline{x}(f) \} \right. & \text{si } f \in s_2 \\ \left. + b_*^{\bar{z}_1}(f) \{ \underline{z} - \underline{z}' \} \right\} & \text{si } f \in s_1 \end{array} \right\}.$$

La différence entre (3.14) et (3.13) peut s'exprimer

$$(3.15) \quad \left. \begin{aligned} & \{ (f) \varepsilon_7^* q^z \} \\ & \{ (f) \varepsilon_7^* q^z + (f) \varepsilon_7^* q^z + (f) p^z + (f) \varepsilon_7^* q^z + (f) \varepsilon_7^* q^z + (f) \varepsilon_7^* q^z \} \end{aligned} \right\} \begin{aligned} & \text{SI } f \ni s_1 - s_2 \\ & \text{SI } f \ni s_2 \end{aligned}$$

variance de l'estimateur  $\hat{Y}_{\text{Chl}}^{\text{Chl}}$  est donné par

$$\begin{aligned} & \left[ (f)_{\tau}^1 \sum_{\tau \ni f}^{\tau \ni f} \{ \tau_* q \} + \right. \\ & \{ (f)_{\tau}^2 g + (f)_{\tau}^1 g \} (f)_{\tau}^2 p (f)_{\tau}^1 q \sum_{\tau \ni f}^{\tau \ni f} \tau_* q \tau + \\ & (f)_{\tau}^1 g (f)_{\tau}^1 g \sum_{\tau \ni f}^{\tau \ni f} \tau_* q \tau + \\ & \left. \{ (f)_{\tau}^1 g \tau + (f)_{\tau}^2 g \} (f)_{\tau}^2 g \sum_{\tau \ni f}^{\tau \ni f} \{ (f)_{\tau}^1 q \} + \right. \\ & \left. (f)_{\tau}^2 p_{\tau} \{ (f)_{\tau}^1 q \} \sum_{\tau \ni f}^{\tau \ni f} + (f)_{\tau}^1 g \sum_{\tau \ni f}^{\tau \ni f} \right] \{ w / (1 - w) \} \\ & = (A^{\text{LCK}})^{\text{LCK}} \end{aligned}$$

#### 4. Etude par simulations

A la présente section, nous présentons les résultats d'études par simulations conçues pour examiner les propriétés de la méthode du jackknife proposée lorsqu'elle est utilisée pour estimer la variance de quatre des estimateurs à deux phases de la moyenne de population présentés à la section 3. Plus précisément, nous considérons l'estimateur de type ratio de Rao et Sitter (1995), l'estimateur de type régression de Sitter (1997), l'estimateur de type ratio en chaîne de Chand (1975) et l'estimateur de type régression en

et  $\hat{X}_o^2(j)$  et  $\hat{X}_o^1(j)$  sont définis de manière analogue. Si  $\hat{R} = \hat{Y}_o^2 / \hat{X}_o^2$  et  $w_{oI}^2 = (1/p_I) / \sum_{i \in s_2} (1/p_i)$ , la différence entre (3.7) et (3.6) peut facilement s'écrire sous la forme

$$\hat{Y}_{\text{Raj}}^c(j) - \hat{Y}_{\text{Raj}}^c =$$

$$\left\{ \begin{array}{l} -w_{oI}^2 \frac{\hat{X}_o^1(j)}{\hat{X}_o^2(j)} (y_j - \hat{R}x_j) \\ + \hat{R} \{ \hat{X}_o^1(j) - \hat{X}_o^1 \} \quad \text{si } j \in s_2 \\ \hat{R} \{ \hat{X}_o^1(j) - \hat{X}_o^1 \} \quad \text{si } j \in (s_1 - s_2). \end{array} \right.$$

Donc, l'estimateur jackknife de la variance de l'estimateur

$$\hat{Y}_{\text{Raj}}^c \text{ est donné par}$$

$$\hat{V}_{\text{Raj}}^c(\hat{Y}_{\text{Raj}}^c) =$$

$$\frac{m-1}{m} \left[ \sum_{j \in s_2} (w_{oI}^2)^2 \frac{\hat{X}_o^1(j)}{\hat{X}_o^2(j)} (y_j - \hat{R}x_j)^2 \right.$$

$$\left. + \hat{R}^2 \sum_{j \in s_1} \{ \hat{X}_o^1(j) - \hat{X}_o^1 \}^2 \right.$$

$$\left. - 2\hat{R} \sum_{j \in s_1} w_{oI}^2 \frac{\hat{X}_o^1(j)}{\hat{X}_o^2(j)} (y_j - \hat{R}x_j) \{ \hat{X}_o^1(j) - \hat{X}_o^1 \} \right].$$

À l'instar de Rao et Sitter (1995), si nous supposons que  $\hat{X}_o^1(j) / \hat{X}_o^2(j) \approx \hat{X}_o^1 / \hat{X}_o^2$ , l'estimateur jackknife de la variance de  $\hat{Y}_{\text{Raj}}^c$  prend la forme

$$\hat{V}_{\text{JACK}}^c(\hat{Y}_{\text{Raj}}^c) \approx$$

$$\frac{m-1}{m} \left[ \sum_{j \in s_2} \{ \hat{X}_o^1 / \hat{X}_o^2 \}^2 \sum_{j \in s_2} (w_{oI}^2)^2 (y_j - \hat{R}x_j)^2 \right.$$

$$\left. + \hat{R}^2 \sum_{j \in s_1} \{ \hat{X}_o^1(j) - \hat{X}_o^1 \}^2 \right.$$

$$\left. - 2\hat{R} \{ \hat{X}_o^1 / \hat{X}_o^2 \} \sum_{j \in s_2} w_{oI}^2 (y_j - \hat{R}x_j) \{ \hat{X}_o^1(j) - \hat{X}_o^1 \} \right].$$

### Cas 3.4 : Srivenkataramana et Tracy (1989)

Afin d'examiner ce cas, comme dans Raj (1965), nous supposons que l'échantillon initial  $s_1$  de taille  $m$  est sélectionné avec remise avec probabilités proportionnelles à  $z_i$ . Cependant, le sous-échantillon,  $s_2$ , de  $n$  unités est maintenant sélectionné avec remise avec probabilités proportionnelles à  $x_i / z_i$ . Par conséquent,  $w_{iI}^1 = (1/z_i) / \sum_{i \in s_1} (1/z_i)$  et  $w_{iI}^2 = (1/x_i) / \sum_{i \in s_2} (1/x_i)$ . Comme dans Raj (1965), nous n'effectuons aucun calage de première phase; donc  $\hat{X}_i^1 = \hat{X}_i^1$ . D'où, si  $q_{2i} = 1/x_i$ , l'estimateur calé  $\hat{Y}_c^{\text{ST}}$  est

$$\hat{Y}_c^{\text{ST}} = \hat{Y}_o^2(\hat{X}_o^1 / \hat{X}_o^2), \quad (3.8)$$

où  $\hat{Y}_o^2 = \sum_{i \in s_2} (y_i/x_i) / \sum_{i \in s_2} (1/x_i)$ ,  $\hat{X}_o^2 = n / \sum_{i \in s_2} (1/x_i)$  et  $\hat{X}_o^1 = \sum_{i \in s_1} (x_i/z_i) / \sum_{i \in s_1} (1/z_i)$ . Donc, alternativement,  $\hat{Y}_c^{\text{ST}} = \sum_{i \in s_2} \{ \hat{Y}_o^2 / \hat{X}_o^2 \} \{ \hat{X}_o^1(j) / \hat{X}_o^2(j) \}$  si  $j \in s_2$  et  $\hat{Y}_c^{\text{ST}}(j) = \hat{Y}_o^2 \{ \hat{X}_o^1(j) / \hat{X}_o^2 \}$  si  $j \in (s_1 - s_2)$

leur jackknife de la moyenne de population est

$$\hat{Y}_c^{\text{ST}}(j) = \begin{cases} \hat{Y}_o^2 \{ \hat{X}_o^1(j) / \hat{X}_o^2(j) \} & \text{si } j \in s_2 \\ \hat{Y}_o^2 \{ \hat{X}_o^1(j) / \hat{X}_o^2 \} & \text{si } j \in (s_1 - s_2) \end{cases} \quad (3.9)$$

où

$$\hat{Y}_o^2(j) = \frac{\sum_{i \in s_2} (y_i/x_i)}{\sum_{i \in s_2} (1/x_i)} + \frac{x_j \sum_{i \in s_2} (1/x_i) - 1}{\sum_{i \in s_2} (1/x_i)} - 1$$

Les termes  $\hat{X}_o^2(j)$  et  $\hat{X}_o^1(j)$  sont définis similairement; autrement dit

$$\hat{X}_o^2(j) = \frac{\sum_{i \in s_2} (1/x_i)}{n} + \frac{x_j \sum_{i \in s_2} (1/x_i) - 1}{\sum_{i \in s_2} (1/x_i)} - 1$$

tandis que  $\hat{X}_o^1(j)$  peut s'écrire

$$\hat{X}_o^1(j) = \frac{\sum_{i \in s_1} (x_i/z_i)}{\sum_{i \in s_1} (1/x_i)} + \frac{x_j \sum_{i \in s_1} (1/x_i) - 1}{\sum_{i \in s_1} (1/x_i)} - 1$$

Si  $\hat{R} = \sum_{i \in s_2} (y_i/x_i) / n$  et  $w_{oI}^2 = (1/x_j) / \sum_{i \in s_2} (1/x_i)$ , la différence entre (3.9) et (3.8) est donnée par

$$\hat{Y}_c^{\text{ST}}(j) - \hat{Y}_{\text{Raj}}^c =$$

$$\left\{ \begin{array}{l} -w_{oI}^2 \frac{\hat{X}_o^1(j)}{\hat{X}_o^2(j)} (y_j - \hat{R}x_j) + \hat{R} \{ \hat{X}_o^1(j) - \hat{X}_o^1 \} \quad \text{si } j \in s_2 \\ \hat{R} \{ \hat{X}_o^1(j) - \hat{X}_o^1 \} \quad \text{si } j \in (s_1 - s_2). \end{array} \right.$$

Suivant Rao et Sitter (1995), si nous supposons que  $\hat{X}_o^1(j) / \hat{X}_o^2(j) \approx \hat{X}_o^1 / \hat{X}_o^2$ , l'estimateur jackknife de la variance de  $\hat{Y}_{\text{Raj}}^c$  prend la forme

$$\hat{V}_{\text{JACK}}^c(\hat{Y}_{\text{Raj}}^c) \approx$$

$$\frac{m-1}{m} \left[ \sum_{j \in s_2} \{ \hat{X}_o^1 / \hat{X}_o^2 \}^2 \sum_{j \in s_2} (w_{oI}^2)^2 (y_j - \hat{R}x_j)^2 \right.$$

$$\left. + \hat{R}^2 \sum_{j \in s_1} \{ \hat{X}_o^1(j) - \hat{X}_o^1 \}^2 \right.$$

$$\left. - 2\hat{R} \{ \hat{X}_o^1 / \hat{X}_o^2 \} \sum_{j \in s_2} w_{oI}^2 (y_j - \hat{R}x_j) \{ \hat{X}_o^1(j) - \hat{X}_o^1 \} \right].$$

### Cas 3.5 : Chand (1975)

Afin d'examiner ce cas, l'échantillon de première phase  $s_1$  de taille  $m$  est sélectionné par EASSR et les variables auxiliaires  $Z$  et  $X$  sont toutes deux observées sur les unités choisies. Le sous-échantillon,  $s_2$ , de  $n$  unités est



où  $\bar{y} = \sum_{i \in s_2} y_i/n$ ,  $\bar{x} = \sum_{i \in s_2} x_i/m$ . Dans

(2.1), le mécanisme jackknife devient

$$\hat{Y}_{RS}^{jc}(f) = \begin{cases} (\bar{ny} - y_j)(m\bar{x}' - x_j) / (n\bar{x} - x_j)(m - 1) & \text{si } j \in s_2 \\ (\bar{x}/x_j)(m\bar{x}' - x_j) / (m - 1) & \text{si } j \in (s_1 - s_2). \end{cases} \quad (3.2)$$

En fixant  $R = \bar{y}/\bar{x}$ , la différence entre (3.2) et (3.1) peut s'écrire

$$\hat{Y}_{RS}^{jc}(f) - \hat{Y}_{RS}^{jc} = \begin{cases} -R \frac{(x_j - \bar{x}')}{(x_j - \bar{x})} \frac{(m - 1)}{(m - 1)} & \text{si } j \in (s_1 - s_2). \\ -R \frac{(x_j - \bar{x}')}{(x_j - \bar{x})} \frac{(m - 1)}{(m - 1)} - \frac{\bar{x}'(j)}{\bar{x}(j)} \frac{(m - 1)}{(y_j - Rx_j)} & \text{si } j \in s_2 \end{cases} \quad (3.3)$$

L'expression (3.3) est exactement la même que celle donnée par Rao et Sitter (1995). En supposant que  $\bar{x}'(j)/\bar{x}(j) \approx x_j'/x_j$ , l'estimateur jackknife approximatif de la variance est donné par

$$\hat{V}_{JACK}^{jc}(\hat{Y}_{RS}^{jc}) \approx \left( \frac{\bar{x}'}{\bar{x}} \right)^2 \sum_{i \in s_2} \frac{(y_i - Rx_i)^2}{n(n - 1)} + 2 \left( \frac{\bar{x}'}{\bar{x}} \right) R \sum_{j \in s_2} \frac{x_j}{(x_j - \bar{x}')(y_j - Rx_j)} \frac{n - 1}{(x_j - \bar{x}')^2} + R^2 \sum_{j \in s_1} \frac{m(m - 1)}{(x_j - \bar{x}')^2}.$$

Donc, l'estimateur de Rao et Sitter (1995) est un cas particulier de la méthode du jackknife proposée.

### Cas 3.2 : Sitter (1997)

Dans le cas 3.1, si nous considérons  $q_{2i} = 1$ , l'estimateur calé sous EASSR devient

$$\hat{Y}_{pc}^{jc} = \bar{y} + b^*(\bar{x}' - \bar{x}), \quad (3.4)$$

où  $b^* = \sum_{i \in s_2} x_i y_i / \sum_{i \in s_2} x_i^2$  désigne un estimateur du coefficient de régression  $\beta$  qui est légèrement différent de celui envisagé par Sitter (1997). Le mécanisme du jackknife prend la forme

$$\hat{Y}_{pc}^{jc}(f) = \begin{cases} \frac{n\bar{y} - y_j}{n - 1} + \left\{ b^* + \frac{x_j(y_j - b^* x_j)}{\sum_{i \in s_2} x_i^2} \right\} \frac{n - 1}{m\bar{x}' - x_j} - \frac{m - 1}{x_j - \bar{x}} & \text{si } j \in s_2 \\ \left\{ \bar{y} + b^* \right\} \frac{m - 1}{m\bar{x}' - x_j} - \frac{m - 1}{x_j - \bar{x}} & \text{si } j \in (s_1 - s_2). \end{cases} \quad (3.5)$$

ou

$$\hat{Y}_{pc}^{jc}(f) = \frac{\sum_{i \in s_2} (y_i/d_i)}{\sum_{i \in s_2} (1/d_i)} + \frac{\sum_{i \in s_2} (1/d_i)}{\sum_{i \in s_2} (1/d_i)} \left\{ 1 - \frac{\sum_{i \in s_2} (1/d_i)}{\sum_{i \in s_2} (1/d_i)} \right\} \frac{\sum_{i \in s_2} (1/d_i)}{\sum_{i \in s_2} (y_i/d_i)} - y_j \quad (3.6)$$

$$\hat{Y}_{pc}^{jc}(f) = \begin{cases} \frac{Y_{pc}^{jc}(f)}{X_{pc}^{jc}(f)} & \text{si } j \in s_2 \\ \frac{Y_{pc}^{jc}(f)}{X_{pc}^{jc}(f)} & \text{si } j \in (s_1 - s_2) \end{cases} \quad (3.7)$$

leur jackknife de la moyenne de population est

Sous le plan d'échantillonnage décrit ci-dessus, l'estimateur calé sous EASSR devient

où  $\hat{Y}_o^2 = \sum_{i \in s_2} (y_i/d_i)/\sum_{i \in s_2} (1/d_i)$ ,  $\hat{X}_o^2 = \sum_{i \in s_2} (x_i/d_i)/\sum_{i \in s_2} (1/d_i)$ . Donc, alternativement,  $\hat{Y}_{raj}^{jc} = \{\sum_{i \in s_2} (y_i/d_i)/\sum_{i \in s_2} (1/d_i)\} / \{\sum_{i \in s_2} (x_i/d_i)/\sum_{i \in s_2} (1/d_i)\}$ .

$$\hat{Y}_{raj}^{jc} = \hat{Y}_o^2 / \hat{X}_o^2, \quad (3.6)$$

Afin d'examiner ce cas, nous supposons que l'échantillon initial  $s_1$  de taille  $m$  est sélectionné avec remise avec les probabilités  $p_i$  proportionnelles à  $z_i$ ,  $i = 1, 2, \dots, N$ . L'information sur la variable auxiliaire  $X$  est recueillie sur cet échantillon de première phase,  $s_1$ . L'échantillon de deuxième phase, dont la taille est fixée à  $n$ , est un sous-échantillon de  $s_1$  sélectionné sans remise avec probabilités égales. C'est pour cet échantillon  $s_2$  que l'information sur  $Y$  est recueillie. Sous ce plan d'échantillonnage,  $d_{1i} = 1/\pi_{1i} = 1/(mp_i)$  et  $d_{2i} = m/n$ . Donc,  $w_i^o = (1/d_i)/\sum_{i \in s_1} (1/d_i)$  et  $w_i^2 = (1/d_i)/\sum_{i \in s_2} (1/d_i)$ . Notons aussi que, pour ce plan,  $\hat{X}_o^1 = \hat{X}_o^2$ ; donc, aucun calage de première phase n'est effectué. Si  $q_{2i} = 1/x_i$ , l'estimateur calé  $\hat{Y}_c^{jc}$  devient

### Cas 3.3 : Raj (1965)

qui est semblable à l'expression donnée par Sitter (1997).

$$\hat{Y}_{pc}^{jc}(f) - \hat{Y}_{pc}^{jc} = \begin{cases} -b^* \frac{(x_j - \bar{x}')}{(x_j - \bar{x})} \frac{(m - 1)}{(m - 1)} & \text{si } j \in (s_1 - s_2) \\ -b^* \frac{(x_j - \bar{x}')}{(x_j - \bar{x})} \frac{(m - 1)}{(m - 1)} - \frac{1 + \frac{a_j^*}{a_j^*}}{\frac{d_j^*}{(n - 1)}} & \text{si } j \in s_2 \end{cases}$$

forme

Si nous posons que  $d_j^* = (y_j - \bar{y}) - b^*(x_j - \bar{x})$ ,  $a_j^* = x_j \{ \bar{x}(j) - \bar{x}'(j) \} / K$  et  $K_j = x_j^2 / K$ , où  $K = (n - 1)s_2^2 + n\bar{x}^2$ , la différence entre (3.5) et (3.4) peut s'écrire sous la

(2000, 2006) en ce sens que nous considérons le calage à la première ainsi qu'à la deuxième phase, ce qui permet d'élaborer la méthode des estimateurs par le ratio en chaîne et par la régression en chaîne. Nous évaluons aussi, dans une étude par simulations, l'efficacité des estimateurs jackknife de variance comparativement aux estimateurs de variance usuels.

## 2. Estimation de la variance par la méthode du jackknife

Dans la suite de l'exposé, nous supposons qu'un plan à un seul degré est employé aux deux phases du processus d'échantillonnage. Soit  $\hat{Y}^c(j)$  un estimateur calé de la moyenne de population,  $\bar{Y}$ , obtenu en retranchant la  $j^e$  unité de l'échantillon  $s_1$  de  $m$  unités. Nous prouvons en annexe que l'estimateur jackknife de la moyenne de population sous échantillonnage à deux phases peut s'écrire

$$\hat{Y}^c(j) = \begin{cases} \hat{Y}^c(j) + \hat{\beta}_2(j) \{X^c_1(j) - X^c_2(j)\} \\ + \hat{\beta}_1(j) \hat{\beta}_2(j) \{\bar{Z} - \bar{Z}^c_1(j)\} & \text{si } j \in s_2 \\ \hat{Y}^c_0 + \hat{\beta}_2 \{X^c_1(j) - X^c_0\} \\ + \hat{\beta}_1(j) \hat{\beta}_2 \{\bar{Z} - \bar{Z}^c_1(j)\} & \text{si } j \in (s_1 - s_2) \end{cases} \quad (2.1)$$

où la quantité  $\bar{Z}^c_1(j) = \bar{Z}^c_0 + \{w^c_{1j}/(1 - w^c_{1j})\} \{\bar{Z}^c_1 - z_j\}$ , les termes  $X^c_1(j)$ ,  $X^c_2(j)$  et  $\bar{Y}^c(j)$  sont définis de manière analogues,  $\hat{\beta}_1(j) = \hat{\beta}_1 + \{q_{1j}w^c_{1j}z_j(x_j - \hat{\beta}_1z_j)\} / \{q_{1j}w^c_{1j}z^2_j - \sum_{i \in s_1} q_{1i}w^c_{1i}z^2_i\}$  et  $\hat{\beta}_2(j) = \hat{\beta}_2 + \{q_{2j}w^c_{2j}x_j(y_j - \hat{\beta}_2x_j)\} / \{q_{2j}w^c_{2j}x^2_j - \sum_{i \in s_1} q_{2i}w^c_{2i}x^2_i\}$ . L'estimateur jackknife modifié de la variance de  $\hat{Y}^c$  est alors donné par

$$V^{JACK}(\hat{Y}^c) = \{(m - 1)/m\} \sum_{j \in s_1} \{\hat{Y}^c(j) - \hat{Y}^c\}^2. \quad (2.2)$$

Nous montrons en annexe que cet estimateur est convergent. Notons que nous pouvons écrire

$$\hat{Y}^c(j) - \hat{Y}^c = \begin{cases} \varepsilon_1(j) + \hat{\beta}_2\varepsilon_1(j) + \hat{\beta}_2\varepsilon_2(j)d_2(j) \\ + \hat{\beta}_2\delta_2(j) & \text{si } j \in s_2 \\ \hat{\beta}_2\varepsilon_1(j) & \text{si } j \in (s_1 - s_2) \end{cases}$$

où les termes de (2.3) sont donnés par  $\varepsilon_1(j) = \{X^c_1(j) - \hat{X}^c_1\} - \hat{\beta}_1\{Z^c_1(j) - \bar{Z}\} - \hat{\beta}_2\{X^c_2(j) - \hat{X}^c_2\} - \hat{\beta}_1\{Z^c_2(j) - \bar{Z}\}$  et  $\delta_2(j) = \{X^c_2(j) - \hat{X}^c_2\} - \hat{\beta}_1\{Z^c_2(j) - \bar{Z}\} - \hat{\beta}_2\{X^c_1(j) - \hat{X}^c_1\} - \hat{\beta}_1\{Z^c_1(j) - \bar{Z}\}$ . Le terme  $\varepsilon_1(j)$  est analogue au

En utilisant (2.3) dans (2.2), l'estimateur jackknife de la variance de l'estimateur  $\hat{Y}^c$  est donné par

$$V^{JACK}(\hat{Y}^c) = \{(m - 1)/m\} \left[ \sum_{j \in s_1} \hat{\beta}^2_2(j)d^2_2(j) + \sum_{j \in s_2} \hat{\beta}^2_2(j)\{\delta_2(j) + 2\varepsilon_1(j)\} + 2\hat{\beta}_2 \sum_{j \in s_2} \varepsilon_1(j)\varepsilon_2(j) + 2\hat{\beta}_2 \sum_{j \in s_2} \hat{\beta}_2(j)d_2(j)\{\varepsilon_1(j) + \delta_2(j)\} + \hat{\beta}^2_2 \sum_{j \in s_1} \varepsilon^2_2(j) \right]. \quad (2.4)$$

Notons que l'expression (2.4) est exacte. Elle peut être utilisée pour estimer la variance de plusieurs estimateurs décrits dans la littérature.

### 3. Cas particuliers

À la section suivante, nous démontrons que les estimateurs proposés par Rao et Sitter (1995), Sitter (1997), Raj (1965), Srivenukataramana et Tracy (1989), Chand (1975) et Ahmed (1997) peuvent être considérés comme des cas particuliers de la méthode proposée.

#### Cas 3.1 : Rao et Sitter (1995)

Si  $\hat{X}^c_1 = \hat{X}^c_0$  (aucun calage de première phase n'est effectué) et  $q_{2i} = 1/x_i$ , l'estimateur calé de  $\bar{Y}$  devient

$$\hat{Y}^c = \left( \sum_{j \in s_2} w^c_{2j}y_j \right) / \left( \sum_{j \in s_2} w^c_{2j}x_j \right).$$

Si l'échantillon de première phase  $s_1$  est sélectionné selon un plan EAASSR tel que les poids de sondage de première phase sont donnés par  $d_{1i} = N/m$  et que l'échantillon de deuxième phase  $s_2$  est sélectionné à partir de  $s_1$  par BAASSR de façon que  $d_{2i} = m/n$ , l'estimateur calé de la moyenne de population devient

$$\hat{Y}^{RS} = \bar{y}(x'/x), \quad (3.1)$$

méthode du jackknife est plus commode et efficace que par les méthodes classiques fondées sur les estimations des moments.

Dernièrement, un certain nombre d'auteurs ont étudié l'emploi de procédures jackknife pour estimer les variances (voir Arnab et Singh 2006, Berger 2007, Berger et Skinner 2005, Chen et Shao 2001, et Kovar et Chen 1994). Fuller (1998), Kim, Navarro et Fuller (2000, 2006), Kim et Sitter (2003), ainsi que Kott et Stukel (1997) ont proposé une approche d'estimation de la variance dans le cas de l'échantillonnage à deux degrés. Fuller (1998), ainsi que Kim et Sitter (2003) se sont penchés sur l'estimateur par la régression. En particulier, ils ont considéré l'estimateur par la régression généralisée du total de population

$$Y^{DS} = \sum_{i \in s_1} \alpha_i y_i$$

du à Deville et Särndal (1992). Selon Kim et coll. (2000, 2006), pour chaque  $k \in s_2$ , l'estimateur jackknife du total de population est spécifié comme étant

$$Y^{kim} = \sum_{i \in s_2 \setminus k} \alpha_i y_i \quad (1.7)$$

et la distance du khi-deux entre les poids de sondage et les poids calés comme étant

$$D^{(k)} = (1/2) \sum_{i \in s_2 \setminus k} \{(\alpha_i^{(k)} - w_{*(k)}^i w_{(k)}^i)^2 / (w_{(k)}^i q_{(k)}^i)\}. \quad (1.8)$$

La minimisation de (1.8) sous la contrainte

$$\sum_{i \in s_1} \alpha_i x_i = \sum_{i \in s_1 \setminus k} w_{(k)}^i x_i$$

mène aux poids calés jackknife donnés par

$$\alpha_i^{(k)} = w_{(k)}^i w_{*(k)}^i + \left\{ (w_{(k)}^i q_{(k)}^i x_i) / \left( \sum_{i \in s_2 \setminus k} w_{(k)}^i q_{(k)}^i \right) \right\}$$

$$\left\{ \sum_{i \in s_2 \setminus k} w_{(k)}^i x_i - \sum_{i \in s_2 \setminus k} w_{(k)}^i w_{*(k)}^i x_i \right\}.$$

Il semble que Kim et coll. (2006) ont rajusté ces poids de la façon suivante

$$\alpha_i^{(k)} = \begin{cases} w_{(k)}^i & \text{si } i \in s_1 \\ w_{(k)}^i & \text{si } i \in s_2 \end{cases}$$

Pour un tel rajustement, l'estimateur donné par (1.7) est équivalent à celui de Rao et Sitter (1995).

Dans le présent article, nous considérons une nouvelle méthode jackknife pour estimer la variance de l'estimateur  $\bar{Y}^c$  sous échantillonnage à deux phases en suivant Hidroglou et Särndal (1995, 1998). Comme l'estimateur de Rao et Sitter (1995) est un cas particulier de la méthode proposée. Cependant, notre approche diffère de celle de Fuller (1998), de Kim et Sitter (2003) et de Kim et coll.

Maintenant, désignons par  $w_{2i}^o = d_{1i} d_{2i} / \sum_{i \in s_2} d_{1i} d_{2i}$  les poids de sondage normalisés de deuxième phase. L'estimateur habituel de  $\bar{Y}$  est donné par

$$\hat{Y}_o^2 = \sum_{i \in s_2} w_{2i}^o y_i.$$

Considérons l'estimateur de deuxième phase calé de  $\bar{Y}$  de la forme

$$\hat{Y}^c = \sum_{i \in s_2} w_{2i}^c y_i \quad (1.3)$$

où les  $w_{2i}^c$  sont les poids calés de deuxième phase tels que la fonction de distance du khi-deux

$$D_2 = \sum_{i \in s_2} \{(w_{2i}^c - w_{2i}^o)^2 / (w_{2i}^o q_{2i}^o)\} \quad (1.4)$$

est minimisée sous la contrainte de calage

$$\sum_{i \in s_2} w_{2i}^c x_i = \hat{X}_1^c. \quad (1.5)$$

La minimisation de (1.4) sous la contrainte (1.5) donne les poids de deuxième phase calés

$$w_{2i}^c = w_{2i}^o$$

$$+ \left\{ (q_{2i} w_{2i}^o x_i) / \left( \sum_{i \in s_2} q_{2i} w_{2i}^o x_i^2 \right) \right\} \left( \hat{X}_1^c - \sum_{i \in s_2} w_{2i}^o x_i \right).$$

Donc, l'estimateur de deuxième phase calé de  $\bar{Y}$  spécifié en

(1.3) peut s'écrire

$$\hat{Y}^c = \hat{Y}_o^2 + \hat{\beta}_2 (\hat{X}_1^c - \hat{X}_o^2) + \hat{\beta}_1 \hat{\beta}_2 (\bar{Z} - \bar{Z}_o^2), \quad (1.6)$$

$$\text{où } \bar{Z}_o^2 = \sum_{i \in s_1} w_{2i}^o z_i, \quad \hat{X}_o^2 = \sum_{i \in s_1} w_{2i}^o x_i, \quad \hat{X}_2^2 = \sum_{i \in s_2} w_{2i}^o x_i, \quad \text{et } \hat{Y}_o^2 = \sum_{i \in s_2} w_{2i}^o y_i,$$

Hidroglou et Särndal (1995, 1998), ainsi que Singh (2000) ont étudié le problème de l'estimation de la variance de l'estimateur calé  $\hat{Y}^c$  donné par (1.6) en utilisant une approche fondée sur le plan de sondage. Dans un contexte plus général, Rao et Sitter (1995), ainsi que Sitter (1997) ont fait observer que, sous échantillonnage aléatoire simple sans remise (EASSR), une méthode jackknife peut être utilisée pour estimer les variances des estimateurs par le ratio et par la régression d'une moyenne de population. Ces auteurs ont également signalé que l'estimation de la variance par la



# Certaines contributions à la méthode du jackknife appliquée aux estimateurs sous échantillonnage à deux phases

Patrick J. Farrell et Sarjinder Singh<sup>1</sup>

## Résumé

Dans le présent article, le problème de l'estimation de la variance de divers estimateurs de la moyenne de population sous échantillonnage à deux phases est traité par application de la méthode du jackknife aux poids calés en deux phases de Hidiroglou et Sæmdal (1995, 1998). Nous montrons que plusieurs estimateurs de la moyenne de population décrits dans la littérature sont des cas particuliers de la méthode élaborée ici, y compris ceux proposés par Rao et Sitter (1995) et par Sitter (1997). En nous inspirant de Raj (1965) et de Srivenkataratnam et Tracy (1989), nous introduisons de nouveaux estimateurs de la moyenne de population et nous estimons leur variance par la méthode du jackknife proposée. Nous estimons également la variance des estimateurs en chaîne par le ratio et par la régression dus à Chand (1975) en utilisant le jackknife. Une étude par simulations nous permet d'évaluer l'efficacité des estimateurs jackknife proposés comparativement aux estimateurs de variance usuels.

Mots clés : Information auxiliaire ; calage ; estimation de la moyenne et de la variance ; jackknife ; échantillonnage à deux phases.

## 1. Introduction

Hidiroglou et Sæmdal (1995, 1998) ont fait remarquer que le recours à l'échantillonnage à deux phases pour estimer les caractéristiques d'une population finie est une technique puissante et rentable qui joue donc un rôle très important dans l'échantillonnage. L'échantillonnage à deux phases peut être décrit comme il suit. Considérons une population finie que nous désignons par  $\Omega = \{1, 2, \dots, N\}$ . Supposons que l'information sur une variable  $Z$  est disponible pour l'ensemble de la population ; autrement dit, les valeurs  $Z_i$  pour tout  $i = 1, \dots, N$  sont connues, ce qui implique que la moyenne de population,  $\bar{Z}$ , est également connue. Un échantillon probabiliste de première phase  $s_1, s_1 \subset \Omega$ , de taille  $m$  est tiré de la population avec les probabilités de sélection  $\pi_{i1}$ . Donc, les poids d'échantillonnage de première phase peuvent être définis comme  $d_{i1} = 1/\pi_{i1}$ . Supposons que, pour cet échantillon, l'information est recueillie sur une variable  $X$ , qui est ensuite appariée avec l'information sur  $Z$  pour chacune des  $m$  unités, ce qui fournit les données  $\{(x_i, z_i) \mid i \in s_1\}$  pour  $i = 1, \dots, m$ . Lorsque l'échantillon de première phase  $s_1$  a été tiré, un échantillon de deuxième phase  $s_2, s_2 \subset s_1 \subset \Omega$ , de taille  $n$  est sélectionné à partir de  $s_1$  avec les probabilités de sélection  $\pi_{i2} = \pi_{i1s_1}$ , ce qui permet de définir les poids d'échantillonnage de deuxième phase par  $d_{i2} = 1/\pi_{i2}$ . Dans l'échantillon de deuxième phase, l'information est maintenant recueillie sur une variable  $Y$  pour chaque unité sélectionnée. Cette information est liée à celle disponible antérieurement sur  $Z$  et  $X$  pour ces unités, ce qui fournit les données  $\{(x_i, y_i, z_i) \mid i \in s_2\}$  pour

$i = 1, \dots, n$ . Supposons que nous souhaitons estimer la moyenne de population  $\bar{X}$ , ainsi que la variance de l'estimateur employé. Soit  $w_{i1} = d_{i1}/\sum_{i \in s_1} d_{i1}$  les poids de sondage originaux normalisés de première phase. L'estimateur habituel de la moyenne de population  $\bar{X}$  est donné par

$$\hat{X}_o = \sum_{i \in s_1} w_{i1} x_i$$

tandis qu'un estimateur calé de première phase de  $\bar{X}$  est donné par

$$\hat{X}_c = \sum_{i \in s_1} w_{i1}^* x_i$$

où les  $w_{i1}^*$  sont les poids calés de telle façon que la fonction de distance du khi-deux

$$D_1 = \sum_{i \in s_1} \{(w_{i1}^* - w_{i1})^2 / (w_{i1}^* q_{i1})\} \quad (1.1)$$

est minimisée sous la contrainte

$$\sum_{i \in s_1} w_{i1}^* z_i = \bar{Z}. \quad (1.2)$$

Dans (1.1), les  $q_{i1}$  sont un ensemble de poids choisis de manière appropriée. La minimisation de (1.1) sous la contrainte (1.2) produit les poids calés de première phase

$$w_{i1}^* = w_{i1} + \left\{ (q_{i1} w_{i1} z_i) / \left( \sum_{i \in s_1} q_{i1} w_{i1} z_i^2 \right) \left( \bar{Z} - \sum_{i \in s_1} w_{i1} z_i \right) \right\}.$$

Donc, un estimateur de première phase calé de  $\bar{X}$  est donné par

1. Patrick J. Farrell, School of Mathematics and Statistics, Carleton University, 1125 promenade Colonel By, Ottawa (Ontario), Canada, K1S 5B6. Courriel : pfarrell@math.carleton.ca ; Sarjinder Singh, Department of Mathematics, Texas A&M University - Kingsville, Kingsville, Texas, E-U, 78363. Courriel : sarjinder@yahoo.com.



- Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9(5), 1010-1019.
- Krotki, K. (2007). Combining RDP and Web Panel Surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association* (sous presse).
- Lessler, J.T., et Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. New York : John Wiley & Sons, Inc.
- Lumley, T. (2009). Survey: Analysis of complex survey samples. R package version 3.19. University of Washington : Seattle.
- Mitza, H., et Hömgren, J. (2002). The Sampling and the Estimation Procedure in the Swedish Labour Force Survey. Rapport technique, Statistics Sweden, Stockholm : Sweden.
- Nadimpalli, V., Judkins, D. et Chu, A. (2004). Survey Calibration to CPS Household Statistics. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 4090-4094.
- R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. Disponible au : <http://www.R-project.org>.
- Rao, J.N.K., et Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86(2), 403-415.
- Rao, J.N.K., et Wu, C.F.J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80(391), 620-630.
- Research Triangle Institute (2008). *SUDAAN Language Manual*. Release 10.0, Research Triangle Park, NC : Research Triangle Institute.
- Rust, K.F., et Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- Sämdal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. England : John Wiley & Sons, Inc.
- Sämdal, C.-E., Swensson, B. et Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3), 527-537.
- Sämdal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag, Inc.
- SAS Institute Inc. (2009). *SAS/STAT® 9.2 User's Guide*. Cary, NC : SAS Institute Inc.
- Seate, S.R. (1982). *Matrix Algebra Useful for Statistics*. New York : John Wiley & Sons, Inc.
- StataCorp (2010). *Stata Statistical Software: Release 11*. Survey Data, College Station, TX : StataCorp LP.
- Stukel, D.M., Hidiroglou, M.A. et Sämdal, C.-E. (1996). Estimation de la variance des estimateurs de calage : comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 2, 117-126.
- Taylor, M.F., Brice, J., Buck, N. et Prentice-Lane, E. (2007). *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. University of Essex, Colchester.
- Terhanian, G., Bremer, J., Smith, R. et Thomas, R. (2000). *Correcting Data from Online Survey for the Effects of Nonrandom Selection and Nonrandom Assignment*. Papier de recherche : Harris Interactive.
- Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94(446), 635-644.
- Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.
- Wolter, K.M. (2007). *Introduction to Variance Estimation*. New York : Springer Science+Business Media, LLC.
- Wu, C.F.J. (1985). Variance estimation for the combined ratio and combined regression estimators. *Journal of the Royal Statistical Society, Series B*, 47(1), 147-154.
- Yung, W., et Rao, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 23-31.
- Yung, W., et Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95(451), 903-915.



## Annexe B

## Bibliographie

## Évaluation de l'estimateur MVCE

Pour les calculs qui suivent, soit  $E_B$  et  $\text{Var}_B$  l'espérance et la variance par rapport au plan d'échantillonnage de l'enquête repère. En outre, soit  $E_\varepsilon$  et  $\text{Var}_\varepsilon$  l'espérance et la variance par rapport à la loi normale multivariée à  $G$  dimensions,  $\text{NMV}_G(\mathbf{0}, \mathbf{V}_B)$ . Tous les autres termes sont définis dans le corps de l'exposé.

B.1 : Calcul de  $E[\text{var}^{\text{MVCE}}(\mathbf{N}_B)]$  donnée dans (15)

En utilisant l'expression (14) et  $c_h^2 = m_{Ah}/(m_{Ah} - 1)$ ,

$$E[\text{var}^{\text{MVCE}}(\mathbf{N}_B)] = E_B \left[ E_\varepsilon \left( \frac{\sum_{h=1}^H (m_{Ah} - 1)}{m_{Ah}} \right) \right]$$

$$= \frac{1}{H} \sum_{h=1}^H (\mathbf{N}_B^{(r)} - \mathbf{N}_B)(\mathbf{N}_B^{(r)} - \mathbf{N}_B)' \mathbf{B} \Bigg|_B \Bigg]$$

$$= \frac{1}{H} E_B \left[ \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_\varepsilon(\hat{\varepsilon}^{(r)} | B) \right]$$

$$= \frac{1}{H} \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_B(\mathbf{V}_B) = E_B(\mathbf{V}_B).$$

B.2 : Calcul de  $\text{Var}[\text{var}^{\text{MVCE}}(\mathbf{N}_B)]$  donnée dans (15)

Quand  $y_k = 1$  de sorte que  $\hat{y}^p = \mathbf{1}' \mathbf{N}_B$ ,  $\text{var}^{\text{MVCE}}(\mathbf{1}' \mathbf{N}_B) = H^{-1} \sum_{h=1}^H m_{Ah}^{-1} \sum_{r=1}^{m_{Ah}} \mathbf{1}' \hat{\varepsilon}^{(r)} \hat{\varepsilon}^{(r)} \mathbf{1}$ . En utilisant la formule pour la variance d'une forme quadratique (Searle 1982, section 13.5), nous avons

$$\text{Var}[\text{var}^{\text{MVCE}}(\mathbf{1}' \mathbf{N}_B)]$$

$$= \text{Var}_B \left[ \frac{1}{H} \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_\varepsilon(\mathbf{1}' \hat{\varepsilon}^{(r)} | B) \right] + E_B \left[ \frac{1}{H} \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} \text{Var}_\varepsilon(\mathbf{1}' \hat{\varepsilon}^{(r)} | B) \right]$$

$$= \text{Var}_B \left[ \frac{1}{H} \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} \mathbf{1}' \mathbf{V}_B \mathbf{1} \right]$$

$$+ E_B \left[ \frac{1}{H} \sum_{h=1}^H \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} \{2\mathbf{1}'(\mathbf{1}' \mathbf{V}_B \mathbf{1}) \mathbf{V}_B\} \right]$$

$$= \text{Var}_B[\mathbf{1}' \mathbf{V}_B \mathbf{1}] + \frac{H m_A^*}{2} [E_B(\mathbf{1}' \mathbf{V}_B \mathbf{1})^2],$$

où  $\underline{m}_A^* = (H^{-1} \sum_{h=1}^H m_{Ah}^{-1})^{-1}$  est la moyenne harmonique de  $m_{Ah}$ .

- Bindner, D.A. (1995). Méthodes de linéarisation pour les échantillons à une et deux phases : une approche de type « recette ». *Techniques d'enquête*, 22, 1, 17-22.
- Bray, R., Hourani, L., Rae, K., Dever, J., Brown, J., Vincus, A., Pemberton, M., Marsden, M., Faulkner, D., et Vandermaas-Peeler, R. (2003). 2002 Department of Defense Survey of Health Related Behaviors Among Military Personnel. Rapport technique RTI/7841/006-FR, U.S. Department of Defense préparé par RTI International. URL <http://dodwvs.rti.org/2002WWFfinalReportComplete05-04.pdf>.
- Canty, A.J., et Davison, A.C. (1999). Resampling-based variance estimation for Labour Force Surveys. *The Statistician*, 48, 379-391.
- Centers for Disease Control and Prevention (2006). Technical Information and Data for the Behavioral Risk Factor Surveillance System (BRFSS) – BRFSS Weighing Formula. Atlanta, Georgia : U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 11 septembre 2006.
- Dermati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 1, 17-27.
- Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.
- Deville, J.-C., Särndal, C.-E., et Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88(423), 1013-1020.
- Estévez, V.M., et Särndal, C.-E. (2000). A Functional form approach to calibration. *Journal of Official Statistics*, 16(4), 379-399.
- Fuller, W.A. (1998). Replication variance estimation for the two-phase samples. *Statistica Sinica*, 8, 1153-1164.
- Hidiroglou, M.A., et Pataak, Z. (2006). Raking ratio estimation: An application to the Canadian Retail Trade Survey. *Journal of Official Statistics*, 22(1), 71-80.
- Isaki, C.T., Tsay, J.H. et Fuller, W.A. (2004). Pondération de données d'échantillon reposant sur des contrôles indépendants. *Techniques d'enquête*, 30, 1, 39-49.
- Jayasuriya, B.R., et Valliant, R. (1996). Application de l'estimation par régression restreinte dans une enquête-ménage. *Techniques d'enquête*, 22, 2, 127-138.
- Keefer, S., Dimock, M. et Christian, L. (2008). Calling Cell Phones in '08 Pre-Election Polls. Nouvelle diffusion (18 décembre 2008) : Pew Research Center for the People & the Press. URL <http://people-press.org/reports/pdf/cell-phone-commentary.pdf>.
- Killion, R.A. (2006). Weighing Specifications for The American Time Use Survey (ATUS) for 2006. U.S. Bureau of the Census, Internal Memo (Doc#ATUS-16).
- Kim, J.J., Li, J. et Valliant, R. (2007). Regroupement de cellules lors de la poststratification. *Techniques d'enquête*, 33, 2, 157-170.
- Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 2, 149-160.

### Résultat et la taille relative de l'enquête repère

	Taille relative ( $n_A = 1\ 000$ )	Taille relative ( $n_A = 2\ 000$ )
Variable de résultat	Estimateur de la variance	
NOTCOV	F2CE MVCe	12,0 21,2
PDMED12M	F2CE MVCe	7,7 11,5
		3,8 4,0
		1,1 0,9
		0,4 0,5
		12,0 22,6
		6,3 7,6
		2,1 2,2
		0,7 1,1

## 7. Conclusion et futurs travaux

Les travaux théoriques et analytiques exposés dans le présent article appuient l'idée qu'une nouvelle méthodologie est nécessaire pour traiter la poststratification lorsque l'on utilise des totaux de contrôle estimés, c'est-à-dire la poststratification en fonction de totaux de contrôle estimés (CE). Les estimateurs classiques de la variance peuvent sous-estimer fortement la variance d'échantillonnage de la population, ce qui risque, par exemple, de donner lieu à des décisions incorrectes pour les tests d'hypothèse et à des répartitions non optimales de l'échantillon quand le plan de sondage est mis en œuvre ultérieurement.

L'estimateur de la variance par linéarisation CE var<sup>stce</sup> donne par l'expression (7) est promoteur pour la post-stratification CE. Cet estimateur réduit particulièrement bien le biais relatif en pourcentage observé pour l'estimateur naïf de la variance donné par (6) quand l'enquête repère est petite comparativement à l'enquête analytique. L'estimateur de la variance par répliques var<sup>pcce</sup> donné par (9) est recommandé spécifiquement pour des études nécessitant des poids de répliques, par exemple quand les fichiers d'analyse à grande diffusion sont diffusés sans information sur le plan de sondage pour accroître la protection des données confidentielles de la vie privée des répondants. L'estimateur par répliques de rechange var<sup>mce</sup> donne aussi de bons résultats et est un peu plus facile à appliquer que var<sup>pcce</sup>.

La mise en œuvre des estimateurs de la variance recommandés requiert des programmes informatiques spécialisés, parce que les fonctions requises ne sont pas disponibles à l'heure actuelle dans les logiciels standard. L'estimateur par linéarisation pourrait être plus approchable, parce que sa mise en œuvre comporte une modification en fonction des estimations de la variance disponibles, par exemple  $\text{var}_{\text{STCE}}(\hat{f}^{\text{yPSCe}}) = \text{var}_{\text{Nat}}(\hat{f}^{\text{yPSCe}}) + \sum_A \hat{V}_A^A \sum_B \hat{V}_B^B$ . Nous donnons une discussion étape par étape des procédures requises pour l'estimateur  $\text{var}_{\text{FZCE}}$  (voir la section 4.3) pour faciliter la création du programme informatique.

Des extensions des présents travaux de recherche qui seront présentées à une date ultérieure comprennent une généralisation au calage linéaire, à d'autres statistiques, y

## Remerciements

compris la moyenne estimée par le ratio et à l'estimation par domaine. Nous cherchons aussi à savoir s'il est possible de décaler des valeurs seuils qui déterminent *i)* quand les différences entre les estimations classiques et CE de la variance sont négligeables et *ii)* quand les totaux de contrôle repères sont trop imprécis pour être utilisés pour le calage. Nous prévoyons aussi d'étudier les incidences théoriques des erreurs de mesure dans les enquêtes analytiques ainsi que dans les enquêtes repères.

## Annexe A

### Calcul de $\text{var}_{\text{NJCE}}(\mathbf{N}_B)$

Pour les calculs qui suivent, soit  $E_c$  l'espérance par rapport à une loi normale standard. Tous les autres termes sont définis dans le corps de l'exposé.

$$\text{var}_{\text{NJCE}}(\tilde{\mathbf{N}}^B) = \left( \tilde{\mathbf{N}}^{B(\cdot)} - \frac{1}{H} \sum_{H=1}^H \tilde{\mathbf{N}}^{B(\cdot)} \right)^T \mathbf{K} \left( \frac{1}{M} \sum_{M=1}^M \tilde{\mathbf{N}}^{B(\cdot)} - \frac{1}{H} \sum_{H=1}^H \tilde{\mathbf{N}}^{B(\cdot)} \right)$$

où  $K^{(r)} = \eta^{(r)} \eta'^{(r)}$ , une matrice produit vectoriel de dimension  $\hat{G} \times G$  des valeurs normales et  $\hat{S}_2 = \text{diag}(\hat{V})$ . Parce que  $E^e(K^{(r)}) = I_G$ , une matrice identité de dimension  $G$ , nous avons  $E^e[\text{var}_{N_{JCC}}(\hat{N})] = \text{diag}(\hat{V})$ . D'où,  $\text{var}_{N_{JCC}}(\hat{N}_B)$  ne reproduit pas l'espérance de  $\hat{V}_B$ .



Estimations du biais relatif en pourcentage pour cinq estimateurs de la variance selon la variable de résultat et la taille relative de

L'enquête repère par rapport à l'enquête analytique

Variable de résultat	Estimateur de la variance	Taille relative ( $n_A = 1\ 000$ )	Taille relative ( $n_A = 2\ 000$ )
NOTCOV	Naïf	-50,3	-14,2
	STCE	-4,5	-8,2
	F2CE	-4,7	-8,3
	NJCE	-36,7	-11,9
	MJCE	-4,3	-8,1
	Naïf	-34,4	-10,1
PDMED12M	Naïf	-14,5	-5,7
	STCE	-3,7	-2,7
	F2CE	-3,5	-2,4
	NJCE	-10,5	-4,0
	MJCE	-3,3	-2,4
	Naïf	-34,4	-10,1
	Naïf	-23,4	-10
	STCE	-6,4	-5,1
	F2CE	-6,8	-5,2
	NJCE	-17,6	-7,6
	MJCE	-6,3	-5,0
	Naïf	-48,1	-7,7
	Naïf	-9,2	3,0
	STCE	-56,0	10,8
	F2CE	-0,1	
	NJCE	-40	
	MJCE	-0,2	
	Naïf	-21,7	

biais relatifs plus grands dans ces estimations que ceux produits sous  $n^* = 1\,000$ , même si la taille de l'échantillon de l'enquête analytique est plus grand.

Les tendances qui se dégagent pour le biais relatif en

pourcentage sont reflétées par les *taux de couverture des intervalles de confiance* à 95 % pour les totaux estimés, mais qui ne sont pas présentés par souci de concision. Les estimateurs naïf et NJCE sont plus susceptibles de donner lieu à des taux de couverture des intervalles de confiance inférieurs à 95 %. Ces taux s'approchent du niveau approprié à mesure que la précision des estimations fondées sur l'enquête repère s'améliore. Toutefois, les autres estimateurs CE de la variance ont des taux de couverture proches des niveaux acceptables quelle que soit la taille relative des enquêtes et, par conséquent, sont plus robustes. Jusqu'à présent, la discussion donne à penser que les

différences théoriques, ainsi qu'empiriques, entre les méthodes STCE, F2CE et MVCE sont minimales. Enfin, nous examinons l'écart-type des erreurs-types ( $e_{-t}$ ) estimées pour tenter de distinguer les estimateurs. Un examen de cette variabilité peut donner une idée de la stabilité (empirique) des estimateurs de la variance, c'est-à-dire qu'un estimateur de la variance instable pourrait produire une estimation de la variance médiocre en fonction des nuances d'un échantillon particulier. Le tableau 2 donne l'accroissement relatif en pourcentage des écarts-types pour les méthodes F2CE et MVCE, toutes deux comparativement à la méthode STCE.

La variation des estimations MCVF de la variance est appréciablement plus grande que celle produite par la méthode F2CE, mais uniquement pour des enquêtes de référence relativement petites. L'écart augmente à mesure que la taille de l'échantillon de l'enquête analytique augmente. Ce qui laisse entendre qu'on pourrait préférer la méthode F2CE à la méthode MCVF étant donné la stabilité accrue des estimations de la variance. Toutefois, nous pourrions ces travaux en vue de déterminer le seuil auquel l'instabilité peut avoir une incidence sur les estimations.

Comme prévu, l'estimateur poststratifié classique (naïf)

est celui dont le biais est le plus négatif parmi les estimateurs comparés. Quand l'enquête repère est de plus petite portée que l'enquête analytique (et par conséquent produit des estimations moins précises que cette dernière), l'estimateur natif présente un biais négatif allant jusqu'à 56 %. Le niveau de biais s'améliore à mesure que la taille relative de l'enquête repère augmente ; toutefois, l'estimateur natif produit encore, au mieux, une sous-estimation de 4 %. L'estimateur NJCE donne d'un peu meilleurs résultats que l'estimateur natif, quoique le biais (-2,7 à -40 %) demeure plus grand que pour les autres estimateurs CE de la variance, pour lesquels il varie de -10,1 à 0,1 %.

Pour une petite enquête repère relativement à la taille de

l'enquête analytique (c'est-à-dire la taille relative moins 1), les niveaux de biais (absolus) augmentent spectaculairement pour les estimateurs naïf et NJCCE. Nous constatons l'effet opposé pour les autres estimateurs CE de la variance. La composante de la variance associée à l'enquête repère, c'est-à-dire  $\hat{V}_A^A \hat{V}_B^B \hat{Y}_A^A$  donnée pour  $\text{Var}^{\text{STCE}}$  dans (7), devient le terme dominant dans les estimateurs CE de la variance à mesure que la précision des estimations fondées sur l'enquête repère diminue. Donc, la composante de la variance repère corrige dans une certaine mesure la sous-estimation associée à la composante de la variance analytique. D'autres travaux de recherche sont nécessaires pour déterminer s'il existe un seuil auquel peut avoir lieu ce genre de compensation du biais. Le biais négatif global de nos estimations est comparable au biais des estimateurs de la variance par linéarisation présenté dans un autre contexte par Rao et Wu (1985, section 4) et par Wu (1985). Cependant, les travaux devront se poursuivre afin de déterminer comment réduire au minimum la sous-estimation.

Notons que les tailles relatives de 21,7 quand  $n_A = 1\,000$  et de 10,8 quand  $n_A = 2\,000$  impliquent toutes deux des tailles d'échantillon de l'enquête repère d'environ 21 600. Donc, la composante d'ordre  $O(M^{2/m_B})$  de la variance,  $\hat{\mathbf{V}}_B \hat{\mathbf{V}}_B^A$ , est plus importante pour les estimations du tableau 1 fondées sur  $n_A = 2\,000$ . Cela donne lieu à des



de la base de sondage, ainsi que d'autres problèmes inhérents aux sondages. Chacun des 4 000 échantillons de simulation a été sélectionné de façon à imiter, pour l'enquête analytique, une base de sondage présentant des différences de sous-dénombrement, comme celles utilisées pour de nombreuses enquêtes téléphoniques. Seize ( $G = 16$ ) cellules de post-stratification ont été définies par croisement d'une variable d'âge à huit niveaux avec le sexe. Les taux de couverture pour les 16 cellules ont été créés en se fondant sur les moyennes de population pour chaque groupe d'âge selon le sexe et varient en valeur de 0,5 à 0,9. Un taux de couverture égal à 1 indiquerait une couverture complète. Avant la sélection de chaque échantillon, la base de sondage a été définie comme un sous-échantillon aléatoire stratifié de la population complète de 21 664. Par exemple, 90 % de la population masculine de 65 à 69 ans ont été sélectionnés aléatoirement pour faire partie de la base de sondage pour les simulations de NOTCOV. Ce processus de définition d'un sous-ensemble de la population pour former la base de sondage a été exécuté indépendamment pour chaque échantillon et pour chaque variable de résultat.

Nous soupçonnons que les chercheurs fondent sur la précision des totaux de contrôle leur décision d'utiliser un estimateur de variance par poststratification classique ou par poststratification CE. Nous avons calculé la matrice de covariance repère ( $\mathbf{V}_B$ ) d'après le fichier de données à grande diffusion complet de la NHIS (92 148 enregistrements) et corrigé proportionnellement les valeurs afin de refléter une taille d'échantillon comparable à notre population de simulation ( $N = 21\,664$ ). Les valeurs hors diagonale de  $\mathbf{V}_B$  varient de -0,05 à 0,75 avec une valeur moyenne de 0,22. Partant de cette matrice, nous avons calculé quatre matrices de covariance pour la simulation en divisant la matrice originale par les facteurs de correction 1,0, 3,6, 18 et 72. Les corrections reflètent les enquêtes repères avec une taille effective d'échantillon approximative de 21 700, 6 000 ( $\approx 21\,700/3,6$ ), 1 200 et moins de 500, respectivement.

Nous avons exécuté la simulation au moyen du logiciel R® (Lumley 2009 ; R Development Core Team 2009) étant donné ses très grandes capacités d'analyse de données d'enquête et son efficacité pour les analyses simulées. Nous avons produit un code pour calculer les estimations de la variance par linéarisation et par répliques pour l'estimateur poststratifié CE dont il est question plus haut parce que le code pertinent n'existe pas à l'heure actuelle.

## 5.2 Critère d'évaluation

Nous avons comparé les résultats empiriques pour les cinq estimateurs de variance discutés à la section précédente (naïf, STCE, F2CE, NJCCE et MVCE) en utilisant trois mesures sur l'ensemble des  $j = 1, \dots, 4\,000$  échantillons de nos simulations. Les estimations pour l'estimateur STCE sont un peu plus faibles que celles calculées pour les estimateurs F2CE et MVCE pour des enquêtes repères relativement petites. Cependant, les différences sont négligeables quand la taille de l'enquête repère augmente.

Les biais relatifs en pourcentage produits par notre étude par simulation sont présentés au tableau 1. Les estimations du biais sont plus grandes pour les estimateurs naïf et NJCCE de la variance que pour les autres estimateurs fondés sur des totaux de contrôle estimés (CE) pour toutes nos simulations. Les estimations pour l'estimateur STCE sont un peu plus faibles que celles calculées pour les estimateurs F2CE et MVCE pour des enquêtes repères relativement petites. Cependant, les différences sont négligeables quand la taille de l'enquête repère augmente.

## 6.2 Estimateurs de la variance

Complétant l'évaluation théorique exposée à la section 4, un estimateur de la variance efficace doit produire des résultats empiriques dont le biais relatif en pourcentage est quasi nul ou légèrement positif pour une mesure prudente (voir la section 5.2 pour la formule du biais relatif en pourcentage).

Les biais relatifs en pourcentage produits par notre étude par simulation sont présentés au tableau 1. Les estimations du biais sont plus grandes pour les estimateurs naïf et NJCCE de la variance que pour les autres estimateurs fondés sur des totaux de contrôle estimés (CE) pour toutes nos simulations. Les estimations pour l'estimateur STCE sont un peu plus faibles que celles calculées pour les estimateurs F2CE et MVCE pour des enquêtes repères relativement petites. Cependant, les différences sont négligeables quand la taille de l'enquête repère augmente.

## 6.1 Estimateur ponctuel

### 6. Résultats de l'étude par simulation

Afin de justifier le recours à la poststratification, nous avons d'abord évalué l'estimation d'Horvitz-Thompson ( $\sum_s d^k y^k$ ) pour les deux variables de résultat. Cet estimateur a la réputation d'être sans biais par rapport au plan de sondage dans des conditions parfaites. Le biais relatif en pourcentage indique que l'estimateur HT présente un biais négatif, sous-estimant le total de population de 38 % pour la variable NOTCOV et de 41 % pour la variable PDMED12M. Ces grandes valeurs indiquent qu'une certaine correction est nécessaire pour ces niveaux non négligeables de biais. Pour l'estimateur poststratifié  $\hat{t}_{yp}$ , le biais relatif en pourcentage est nettement plus faible, cet estimateur présentant un biais positif n'excédant pas 2 % pour les deux variables de résultat.

Complétant l'évaluation théorique exposée à la section 4, un estimateur de la variance efficace doit produire des résultats empiriques dont le biais relatif en pourcentage est quasi nul ou légèrement positif pour une mesure prudente (voir la section 5.2 pour la formule du biais relatif en pourcentage).

Les biais relatifs en pourcentage produits par notre étude par simulation sont présentés au tableau 1. Les estimations du biais sont plus grandes pour les estimateurs naïf et NJCCE de la variance que pour les autres estimateurs fondés sur des totaux de contrôle estimés (CE) pour toutes nos simulations. Les estimations pour l'estimateur STCE sont un peu plus faibles que celles calculées pour les estimateurs F2CE et MVCE pour des enquêtes repères relativement petites. Cependant, les différences sont négligeables quand la taille de l'enquête repère augmente.

moyen d'une étude par simulation décrite à la section suivante.

## 5. Description de l'étude par simulation

Nous complétons l'évaluation théorique des cinq estimateurs de variance dont il était question à la section précédente par l'analyse de résultats de simulation.

### 5.1 Paramètres de simulation

La population sur laquelle porte la simulation est un

sous-ensemble aléatoire du fichier à grande diffusion de la National Health Interview Survey (NHIS) de 2003 contenant les enregistrements obtenus pour 21 664 adultes. Nous avons réparti ces enregistrements en 25 strates, contenant chacune six UPE. Nous avons tiré les échantillons dans cette « population » selon un plan d'échantillonnage à deux degrés. Nous avons d'abord sélectionné deux UPE avec remise en utilisant des probabilités proportionnelles au nombre total d'adultes (PPT) dans l'UPE. Dans chaque UPE échantillonnée, nous avons sélectionné des échantillons aléatoires simples de  $(n^{Aht} = )$  20 et 40 personnes sans remise, ce qui a donné des tailles totales d'échantillon de 1 000 et de 2 000, respectivement. Pour notre étude, nous avons considéré deux tailles d'échantillon intra-UPE afin d'évaluer les effets des composantes de la variance d'enquêtes analytiques plus petites, calculées en augmentant  $n_A$  sur la variance de  $t_{yp}$ . Pour chaque combinaison d'échantillons au niveau de l'UPE et au niveau de la personne (c'est-à-dire 50 UPE et soit 1 000 ou 2 000 personnes), nous avons tiré 4 000 échantillons de simulation. Nous avons calculé les estimations des totaux de population et des variances connexes pour deux variables binaires de la NHIS : NOTCOV = 1 indique qu'un adulte n'était pas couvert par une assurance-maladie au cours des 12 mois qui ont précédé l'interview de la NHIS (environ 17 % de la population) et PDMED12M = 1 indique qu'un adulte avait retardé des soins médicaux à cause de leur coût au cours des 12 mois qui ont précédé l'interview (environ 7 % de la population). Nous n'avons pas tenu compte de la non-réponse dans la présente étude par simulation afin de réduire au minimum les facteurs susceptibles d'avoir une incidence sur nos comparaisons. (Nota : Les questions de l'interview sur ces variables figurent dans le questionnaire de base sur la famille au [http://ftp.cdc.gov/pub/Health\\_Statistics/NCHS/Survey\\_Questionnaires/NHIS/2003/qfamlyxx.pdf](http://ftp.cdc.gov/pub/Health_Statistics/NCHS/Survey_Questionnaires/NHIS/2003/qfamlyxx.pdf). Nous nous sommes servis des réponses aux questions FHI.070 et FAU.010/FAU.020 pour générer les variables NOTCOV et PDMED12M, respectivement).

La poststratification peut réduire légèrement les variances. Toutefois, dans les enquêtes-ménages, cette technique est principalement utilisée pour corriger le sous-dénombrement

article. La méthode MCVF utilise la matrice de covariance complète  $\mathbf{V}_B$  et s'appuie sur la théorie des grands échantillons de sorte que les corrections des totaux de contrôle peuvent être modélisées comme étant issues d'une loi normale multivariée (NMV) à  $G$  dimensions. Pour la méthode MCVF, les totaux de contrôle répliqués ont la forme

$$\mathbf{N}_B^{(r)} = \mathbf{N}_B + c_h R_h \hat{\mathbf{e}}^{(r)} \quad (14)$$

où  $\hat{\mathbf{e}}^{(r)}$  est un vecteur de longueur  $G$  de variables aléatoires tel que  $\hat{\mathbf{e}}^{(r)} \sim \text{NMV}_G(\mathbf{0}, \mathbf{V}_B)$ ;  $c_h = \sqrt{m_{Ah}/(m_{Ah} - 1)}$ ; et  $R_h = \sqrt{1/(H m_{Ah})}$ .

L'estimateur de variance jackknife avec suppression d'une unité pour la méthode MCVF se calcule comme il

suit

$$\text{var}_{\text{MCVF}}(\hat{t}_{yp}^{(r)}) = \sum_{h=1}^H \frac{m_{Ah}}{(m_{Ah} - 1)} \sum_{r=1}^{r=1} (\hat{t}_{yp}^{(r)} - \hat{t}_{yp})^2 + c_h R_h \hat{\mathbf{e}}^{(r)} \mathbf{B}_{A^{(r)}}^2, \quad (15)$$

où  $\hat{t}_{yp}^{(r)}$  est calculé comme il est décrit pour la méthode

F2CE en (11), mais avec  $N_{Bg}^{(r)}$  défini par la  $g^{\text{e}}$  composante dans (14). Contrairement à la méthode de Fuller,  $\text{var}_{\text{MCVF}}(\mathbf{N}_B) \neq \mathbf{V}_B$ ; à la place, la méthode MCVF doit s'appuyer sur les propriétés de l'estimateur fondées sur le plan de sondage. L'espérance sous le plan de cet estimateur est évaluée par rapport à la loi NMV conditionnellement aux estimations repères ( $E^c$ ), puis par rapport au plan de l'enquête repère ( $E_B$ ). Comme nous le montrons à l'annexe B.1,

$$E_B[E^c(\text{var}_{\text{MCVF}}(\mathbf{N}_B)|B)] = E_B(\mathbf{V}_B). \quad (16)$$

Si  $\mathbf{V}_B$  est un estimateur approximativement sans biais de  $\mathbf{V}_B$ , la matrice de covariance de population est reproduite en appliquant cette méthode.

Sous la méthode à deux phases de Fuller,  $\text{var}_{\text{F2CE}}(\mathbf{N}_B) = \mathbf{V}_B$  parce que  $\text{var}_{\text{F2CE}}(\mathbf{N}_B) = \mathbf{V}_B$ . Pour comparer d'avantage les méthodes F2CE et MCVF, notons que, si nous définissons  $y_k^k = 1$  dans l'enquête analytique,  $\hat{t}_{yp}^{(r)} = \mathbf{1}' \mathbf{N}_B$ . Comme nous le montrons à l'annexe B.2,

$$\text{var}[\text{var}_{\text{MCVF}}(\mathbf{1}' \mathbf{N}_B)] =$$

$$\text{var}_B[\mathbf{1}' \mathbf{V}_B \mathbf{1}] + \frac{2}{H m_A^*} [E_B(\mathbf{1}' \mathbf{V}_B \mathbf{1})^2] > \text{var}_B[\mathbf{1}' \mathbf{V}_B \mathbf{1}] \quad (17)$$

où  $m_A^*$  est la moyenne harmonique des tailles des échantillons d'UPE par strate dans l'enquête analytique. Cela donne à penser que les espérances en grand échantillon de  $\text{var}_{\text{F2CE}}$  et de  $\text{var}_{\text{MCVF}}$  sont similaires, quoiqu'en pratique que l'estimateur MCVF est vraisemblablement plus variable que l'estimateur F2CE. Nous examinons cette question au



1. Calculer l'estimation en échantillon complet  $\hat{t}_{yp}$  en utilisant l'expression (3).

2. Déterminer les  $G$  valeurs propres  $\lambda_g$  et vecteurs propres  $\mathbf{q}_g$  pour  $\mathbf{V}_{B_g} \sqrt{\lambda_g}$ . Concaténer la matrice  $G \times G$  des  $\mathbf{z}_g$  avec une matrice  $G \times (m_A - G)$  de zéros, et trier aléatoirement les colonnes. Désigner par  $\mathbf{Z}$  cette nouvelle matrice  $G \times m_A$ .

3. Calculer un vecteur de longueur  $m_A$  dont les valeurs sont égales à  $c_h = \sqrt{m_{Ah}/(m_{Ah} - 1)}$ , classé de  $h = 1$  à  $H$ . Peupler chaque ligne d'une matrice  $G \times m_A$ , appelée  $\mathbf{C}$ , avec ce vecteur, c'est-à-dire que les valeurs de ligne sont répliquées. Le vecteur de longueur  $m_A$  des poids de strate jackknife,  $\mathbf{W}_R$ , est créé avec des composantes égales à  $(m_{Ah} - 1)/m_{Ah}$  où l'UPBE supprimée est extraite de la strate  $h$ .

4. Calculer le produit de Hadamard (ou élément par élément) ( $\mathbf{Z} \bullet \mathbf{C}$ . Répliquer le vecteur  $\mathbf{N}_B$  dans les colonnes d'une matrice  $G \times m_A$  et l'ajouter à  $\mathbf{Z} \bullet \mathbf{C}$ . Cette nouvelle matrice  $G \times m_A$ , appelée  $\mathbf{N}_{BR}$ , contient les totaux de contrôle repères répliqués discutés dans l'expression (8) pour chacune des  $m_A$  répliques.

5. Calculer les estimations répliquées  $\hat{y}_{Ag(r)} = \hat{t}_{Ag(r)} / \hat{N}_{Ag(r)}$  en supprimant l'une après l'autre une UPBE du fichier de l'échantillon de l'enquête analytique, en corrigeant les poids pour les UPBE restantes (valeurs  $\mathbf{W}_R$ ) et en sommant les valeurs pondérées pour le numérateur et le dénominateur dans la poststrate  $g$ . Désigner par  $\hat{\mathbf{Y}}_R$  la matrice  $G \times m_A$  résultante.

6. Calculer les  $m_A$  estimations répliquées,  $\hat{t}_{yp(r)}$ , en commençant par multiplier les éléments  $\mathbf{N}_{BR}$  par  $\hat{\mathbf{Y}}_R$ , puis en faisant la somme des valeurs de ligne dans une colonne. Puis, soustraire  $\hat{t}_{yp}$  de chacune des  $m_A$  valeurs et élever les termes au carré, multiplier par les corrections des poids de sous-échantillonnage des UPBE spécifiés en (10) et calculer la somme sur les  $m_A$  estimations. La valeur résultante est la variance estimée en utilisant la méthode de Fuller,  $\text{var}_{F2CE}(\hat{t}_{yp})$ .

#### 4.4 Méthode jackknife de Nadimpalli-Judkins-Chu (NJCE)

Nadimpalli et coll. (2004) ont élaboré un estimateur de variance jackknife avec suppression d'une unité qui perturbe aléatoirement les totaux de contrôle pour l'ensemble complet de répliques au lieu de corriger uniquement un sous-échantillon de répliques comme dans la méthode F2CE. Pour l'enquête repère, les totaux de contrôle répliqués ont la forme suivante :

$$\mathbf{N}^{B(r)} = \mathbf{N}_B + c_h R_h \hat{\mathbf{S}}_B \boldsymbol{\eta}^{(r)} \quad (12)$$

où  $c_h = \sqrt{m_{Ah}/(m_{Ah} - 1)}$ , comme pour la méthode F2CE ;  $R_h = \sqrt{1/(H m_{Ah})}$ , une fonction du nombre total de strates  $(H)$  et d'UPBE ( $m_{Ah}$ ) de l'enquête analytique ;  $\hat{\mathbf{S}}_B$  est une matrice *diagonale* des erreurs-types estimées pour les totaux de contrôle repères ; et  $\boldsymbol{\eta}^{(r)}$  est un vecteur de longueur  $G$  de valeurs générées aléatoirement pour chaque réplique à partir de la loi normale standard. Les autres termes sont spécifiés pour la méthode F2CE après l'expression (8). Notons que les estimations de covariance incluses dans l'estimateur F2CE, c'est-à-dire les valeurs hors diagonales de  $\mathbf{V}_B$ , sont fixées à zéro pour l'estimateur NJCE. L'estimateur de variance jackknife avec suppression d'une unité correspondant du total poststratifié se calcule comme il suit :

$$\text{var}_{\text{NJCE}}(\hat{t}_{yp}) = \sum_{h=1}^H \frac{m_{Ah}}{(m_{Ah} - 1)} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yp(r)} - \hat{t}_{yp})^2 = \sum_{h=1}^H \frac{m_{Ah}}{(m_{Ah} - 1)} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yp(r)} - \hat{t}_{yp})^2$$

$$+ c_h R_h \boldsymbol{\eta}^{(r)T} \hat{\mathbf{S}}_B \mathbf{B}_{A(r)} \boldsymbol{\eta}^{(r)}, \quad (13)$$

où  $\hat{t}_{yp(r)}$  est calculé comme il est décrit pour l'estimateur F2CE en (11), mais avec  $\hat{N}_{B(r)}$  défini par la  $g^e$  composante dans (12). Contrairement à la méthode F2CE, la variance d'échantillon des totaux de contrôle répliqués NJCE donnée en (12) reproduit l'espérance de la matrice de covariance repère  $\mathbf{V}_B$  uniquement si les termes de covariance sont réellement nuls (voir l'annexe A pour des détails). Si  $\mathbf{V}_B$  n'est pas diagonale,  $\text{var}_{\text{NJCE}}$  ne passe pas ce test.

L'utilisation de la méthode NJCE serait plausible dans deux cas : *i)* la matrice de covariance repère complète pour les totaux de contrôle n'est pas disponible (par exemple, estimations tirées d'un rapport précédent) ou *ii)* les termes de covariance sont négatifs de sorte que les valeurs résultantes définies par (12) donneraient lieu à des estimations de variance prudentes. La matrice diagonale pour  $\hat{\mathbf{S}}_B$  serait correcte si les dénombrements de poststrate estimés étaient vraiment non corrélés. Cependant, cette situation est peu probable, à cause de la structure multinomiale de  $\mathbf{N}_B$ . Étant donné les conditions établies pour la méthode NJCE, l'espérance de l'estimateur de variance  $n^e$  approchera pas  $\text{AV}(\hat{t}_{yp})$  donné par (5) ; le terme de biais est relié à la différence entre les espérances sous le plan de  $\hat{\mathbf{S}}_B^2$  et de  $\mathbf{V}_B$ .

#### 4.5 Méthode du jackknife normal multivariée (MVCE)

La méthode normale multivariée (MVCE) est une généralisation de la méthode NJCE qui, autant que nous sachions, est exposée pour la première fois dans le présent



fonction de la variance sous la poststratification classique et d'un terme d'accroissement additif associé à la variation dans les totaux de contrôle, c'est-à-dire  $\text{var}_{\text{STCE}}(\hat{t}^{yp}) = \text{var}_{\text{NAT}}(\hat{t}^{yp}) + \sum_A \hat{V}_A^B \hat{V}_B^A$ .

Idéalement, le fichier d'analyse de l'enquête repère serait

disponible pour calculer les valeurs de  $\hat{V}_B$ . Toutefois, les

chercheurs pourraient devoir se servir d'estimations pu-

blées pour les totaux de contrôle marginaux seulement,

c'est-à-dire des estimations ponctuelles et de variance pour

une seule caractéristique au lieu des dénombrements et des

estimations de covariance pour un ensemble de caracté-

ristiques. Nous discutons plus en détail des incidences

lorsque l'information est limitée à la section 4.4.

#### 4.3 Méthode jackknife à deux phases de Fuller (F2CE)

Isaki, Tsay et Fuller (2004) ont appliqué un estimateur de

variance jackknife par suppression d'une unité à deux

phases élaboré par Fuller (1998) à une situation de post-

stratification CE. Le principe qui sous-tend la méthode de

Fuller (F2CE) consiste à prendre une décomposition

spectrale (valeur propre) de la matrice de covariance repère

$(\hat{V}_B)$ , à construire des corrections des valeurs repères qui

sont une fonction des valeurs propres et des vecteurs propres

résultants, puis à ajouter les corrections au vecteur des

totaux de contrôle repères  $(\hat{N}_B)$  pour créer un ensemble de

répliques des totaux de contrôle. Un sous-ensemble choisi

aléatoirement des  $m_A$  répliques est poststratifié en fonction

des  $G$  répliques des totaux de contrôle construites, où le

nombre total d'UPB doit être égal ou supérieur au nombre

de poststrates, c'est-à-dire  $m_A \geq G$ . Spécifiquement, le

total de contrôle repère pour la  $r^e$  réplique est défini

comme étant

$$\hat{N}_{B(r)} = \hat{N}_B + c_h \hat{Z}'_h \quad (8)$$

où  $\hat{Z}'_h = \delta_{\sum_{g=1}^G \delta_{g(r)}^{(r)}} \delta_{g(r)}^{(r)} \hat{Z}'_g$ ;  $c_h = \sqrt{m_A h / (m_A h - 1)}$ , une

constante reliée à la méthode jackknife avec suppression

d'une unité d'estimation de la variance;  $\delta_{g(r)}$  est un indi-

cateur zéro/un qui identifie les  $G$  (parmi les  $m_A$ ) répliques

choisies aléatoirement pour recevoir une correction;

$\delta_{g(r)} = 1$  si la  $g^e$  composante de la décomposition de la

covariance repère est choisie aléatoirement pour la tâche

sachant que la réplique  $r$  est sélectionnée pour la correction;

et  $\hat{Z}_g = \mathbf{q}_g \sqrt{\lambda_g}$ , une fonction d'un vecteur propre  $(\mathbf{q}_g)$  et

de la valeur propre associée  $(\lambda_g)$  où  $\hat{V}_B = \sum_{g=1}^G \hat{Z}_g \hat{Z}_g'$ , par

définition. Donc, sachant que  $\delta_{g(r)} = 1$  pour une réplique

particulière, un indicateur unique  $\delta_{g(r)}^{(r)}$  doit alors être égal à

un; cependant, si  $\delta_{g(r)} = 0$ , tous les indicateurs  $\delta_{g(r)}^{(r)}$  sont

nuls.

Le jackknife avec suppression d'une unité peut prendre

de multiples formes selon la valeur de centrage. Nous

choisissons l'estimateur de variance un peu prudent centre

autour de l'estimation en échantillon complet pour notre

étude ( $v_4$  dans Wolter 2007, section 4.5). L'estimateur de

variance par le jackknife avec suppression d'une unité,

se calcule comme il suit sous la méthode de

Fuller pour un plan d'échantillonnage stratifié à plusieurs

degrés.

$\text{var}_{\text{F2CE}}(\hat{t}^{yp}) = \sum_{h=1}^H \left( \frac{m_A h}{m_A h - 1} \right) \sum_{r=1}^{r=1} (\hat{t}^{yp(r)} - \hat{t}^{yp})^2$

$= \sum_{h=1}^H \left( \frac{m_A h}{m_A h - 1} \right) \sum_{r=1}^{r=1} (\hat{t}^{yp(r)} - \hat{t}^{yp} + c_h \hat{Z}'_h \hat{B}_A(r)) ^2$  (9)

où les termes de (9) sont définis ci-après. Notons que l'asso-

ciation de la  $r^e$  réplique à une strate particulière du plan de

sondage est définie d'après l'appartenance de l'UPB élimi-

née à la strate. Dans (9), les estimations répétées sont de-

finies comme étant  $\hat{t}^{yp(r)} = \sum_h \sum_{i \in s_h} d_{ih} \delta_{gh}^{(r)} d_{ih}^{(r)}$  et

$\hat{N}_{B(r)} = \sum_h \sum_{i \in s_h} d_{ih} \delta_{gh}^{(r)} d_{ih}^{(r)}$ , où les poids de

sous-échantillonnage des UPB sont calculés comme il suit

$d_{ih}^{(r)} = \begin{cases} 1 & \text{si } r=i, i \in s_{Ah} \\ 0 & \text{si } h \neq h' \text{ pour } r \in s_{Ah} \text{ et } i \in s_{Ah'} \end{cases}$  (10)

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

$d_{ih}^{(r)} = \begin{cases} 1 & \text{si } r=i, i \in s_{Ah} \\ 0 & \text{si } h \neq h' \text{ pour } r \in s_{Ah} \text{ et } i \in s_{Ah'} \end{cases}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

si  $h \neq h'$  pour  $r \in s_{Ah}$  et  $i \in s_{Ah'}$

si  $r=i, i \in s_{Ah}$

contribution du biais (élevé au carré) à l'erreur quadratique moyenne (EQM) totale est faible relativement à la variance.

#### 4. Estimation de la variance pour la PSC

Des estimateurs de variance ont été élaborés pour la poststratification classique et sont disponibles dans les logiciels conçus pour l'analyse des données d'enquête, comme R® (R Development Core Team 2009), SAS® (SAS Institute Inc. 2004), Stata® (StataCorp 2010) et SUDAAN® (Research Triangle Institute 2008). Cependant, peu de travaux portant sur l'estimation de la variance dans le cas de la poststratification en fonction de totaux de contrôle estimés ont été effectués.

Dans les sous-sections qui suivent, nous présentons quatre estimateurs de variance CE pour  $\hat{t}_{yp}$  qui tiennent compte de la variance dans les totaux de contrôle après avoir défini la variance d'échantillonnage de population. Ces estimateurs comprennent un estimateur de variance par linéarisation qui vient d'être développé et trois estimateurs de variance par la méthode du jackknife avec suppression d'une UPE. Dans le cas du jackknife avec suppression d'une UPE, les répliques sont créées en supprimant séquentiellement une UPE et en corrigeant les poids pour les UPE restantes dans la strate correspondante du plan de sondage. Cela donne un total de  $m_A = \sum_{h=1}^H m_{Ah}$  répliques calculé par sommation du nombre d'UPE de l'enquête analytique par strate ( $m^{(h)}$  sur les  $H$  strates ( $h = 1, \dots, H$ )).

Un estimateur de variance efficace reproduira la variance d'échantillonnage de population correspondante espérée. La variance d'échantillonnage de population approximative (ou asymptotique) de  $\hat{t}_{yp} = \mathbf{N}_B' \hat{\mathbf{V}}_A \mathbf{N}_B$  a la forme suivante :

$$AV(\hat{t}_{yp}) = \mathbf{N}_B' \mathbf{V}_A \mathbf{N}_B + 2\mathbf{N}_B' \hat{\mathbf{Cov}}(\mathbf{N}_B, \hat{\mathbf{V}}_A) \mathbf{N}_B + \mathbf{N}_B' \hat{\mathbf{V}}_B \mathbf{N}_B + \mathbf{N}_B' \mathbf{V}_A \mathbf{N}_B + \mathbf{N}_B' \hat{\mathbf{V}}_B \mathbf{N}_B \quad (5)$$

où  $\mathbf{N}_B = E(\mathbf{N}_B)$ , un vecteur de valeurs prévues pour les dénombrements de poststrate repères dans les  $G$  poststrates ;  $\mathbf{N}_B = (\hat{N}_{B1}, \dots, \hat{N}_{BG})'$  est un vecteur de longueur  $G$  de totaux de contrôle exprimés d'après l'enquête repère ;  $\hat{\mathbf{V}}_A$  est un vecteur de longueur  $G$  de composantes de la population de la forme  $\hat{y}_{Ag} = \hat{t}_{ygs} / N_{Ag}$  ;  $\mathbf{V}_A$  est la matrice de (variance-)covariance des composantes estimées du vecteur  $\mathbf{V}_A$  ; et  $\mathbf{V}_B$  est la matrice de covariance des  $G$  estimations de contrôle d'après l'enquête repère  $\mathbf{N}_B$ . La première composante,  $\mathbf{N}_B' \mathbf{V}_A \mathbf{N}_B$ , est la variance approximative de l'estimateur poststratifié classique,  $\hat{t}_{ygs}$ , c'est-à-dire que les estimations repères sont traitées comme si elles étaient fixes. La composante  $\mathbf{N}_B' \mathbf{V}_A \mathbf{N}_B$  est la variance associée aux estimations repères conditionnellement à l'échantillon de l'enquête analytique, c'est-à-dire la composante de la variance de poststratification CE. Parce que nous

supposons que les enquêtes analytiques et repères sont indépendantes, la covariance des estimations d'après les deux enquêtes est, par définition, nulle. Donc, la composante  $\text{Cov}(\mathbf{N}_B, \hat{\mathbf{V}}_A)$  dans (5) est éliminée de l'expression. Krewski et Rao (1981), Rao et Wu (1985) et d'autres ont démontré la convergence asymptotique des estimateurs de variance par linéarisation et par le jackknife pour des fonctions non linéaires. Cependant, cet examen doit être étendu à la poststratification en fonction de totaux de contrôle estimés (CE). Nous discutons de l'ensemble d'estimateurs de variance CE pour la variance d'échantillonnage de la population identifiées plus bas ou élaboré pour notre étude. Nous avons calculé les estimateurs d'échantillon en substituant les estimations d'échantillon aux paramètres de variance correspondants. Nous commençons par évaluer un estimateur de variance poststratifié classique ou naïf qui ne tient pas compte de la variance dans les totaux de contrôle estimés.

#### 4.1 Un estimateur de variance classique pour la poststratification CE (naïf)

Divers estimateurs de variance ont été élaborés pour les estimateurs par poststratification. Dans toutes les méthodes, les totaux de contrôle sont supposés fixes et connus sans erreur. Par conséquent,  $\mathbf{V}_A' \mathbf{V}_B \mathbf{V}_A$ , la deuxième composante (positive) dans l'expression (5), est nulle parce que  $\mathbf{V}_B = \mathbf{0}$  par hypothèse. L'estimateur de variance par linéarisation prend la forme

$$\text{var}_{\text{Naïf}}(\hat{t}_{yp}) = \mathbf{N}_B' \hat{\mathbf{V}}_A \mathbf{N}_B \quad (6)$$

où  $\mathbf{N}_B$  est le vecteur des  $G$  estimations repères des totaux de contrôle, et  $\hat{\mathbf{V}}_A$  est la matrice de covariance estimée des estimations  $\hat{\mathbf{V}}_A = (\hat{t}_{y1}/N_{A1}, \dots, \hat{t}_{yG}/N_{AG})'$ . Comme la deuxième composante dans la deuxième ligne de (5) n'est pas estimée, toute formule de variance élaborée pour la poststratification classique sous-estimera, par définition, la variance d'échantillonnage de la population. Cependant, si les estimations repères sont très précises, la contribution de la composante de variance par poststratification CE à l'estimation globale pourrait être négligeable. Donc, la différence entre les estimations pour la poststratification classique et la poststratification CE sera, dans ces situations, également négligeable.

#### 4.2 Linéarisation par série de Taylor (STCE)

Un estimateur de variance par linéarisation pour  $\hat{t}_{yp}$  prend la forme :

$$\text{var}_{\text{STCE}}(\hat{t}_{yp}) = \mathbf{N}_B' \hat{\mathbf{V}}_A \mathbf{N}_B + \mathbf{N}_B' \hat{\mathbf{V}}_B \mathbf{N}_B \quad (7)$$

où  $\hat{\mathbf{V}}_B$  est la matrice de covariance repère estimée pour l'ensemble des  $G$  totaux de contrôle. Les autres termes sont définis pour l'expression (6). La formule STCE est une



réduction est la plus fructueuse quand les poststrates sont formées de telle manière que soit presque nulle la corrélation intra-poststrate de  $y_k$  avec la probabilité que le  $k^e$  élément soit inclus dans la base de sondage (Kim, Li et Valliant 2007).

Pour évaluer le biais (inconditionnel) par rapport au plan de sondage de  $t_{yp}$ , nous devons tenir compte de la propriété aléatoire de quatre composantes, à savoir les plans de sondage de l'enquête analytique et de l'enquête repère et les proportions des bases de sondage correspondantes à couvrir la population. Suivant les travaux de Kim, Li et Valliant (2007, équation 2), le biais approximatif par rapport au plan de  $t_{yp}$  en tant qu'estimateur du total de population  $t_y = \sum_{k \in U} y_k$  se calcule comme il suit

$$\text{Biais}(t_{yp}) = E(t_{yp}) - t_y \quad \equiv \sum_{g=1}^G \left\{ t_{yg} \left[ \frac{N_g}{N_{Bg}} - 1 \right] + N_{Bg} \text{Cov}(y_g, \phi_{Ag}) \bar{\phi}_{Ag}^{-1} \right\} \quad (4)$$

où  $N_g$  est la taille de population pour l'ensemble d'éléments  $U_g$  dans la poststrate  $g$ ;  $N_{Bg} = E(\hat{N}_{Bg})$ , la valeur prévue des estimations pour la poststrate sous le plan de l'enquête repère;  $\text{Cov}(y_g, \phi_{Ag}) = N_g^{-1} \sum_{k \in U_g} (y_k - \bar{y}_g)(\phi_{Ak} - \bar{\phi}_{Ag})$ , la covariance de population entre la variable de résultat ( $y_k$ ) et les proportions à la couverture ( $\phi_{Ak}$ ) dans la strate  $g$ ;  $\bar{y}_g = t_{yg}/N_g$ , la  $g^e$  moyenne de poststrate de  $y$ ;  $t_{yg} = \sum_{k \in U_g} y_k$ , le total de population de  $y$  dans la poststrate  $g$ ; et  $\bar{\phi}_{Ag} = N_{Ag}/N_g$ , la proportion moyenne à la couverture dans la poststrate sous le plan de l'enquête analytique avec  $N_{Ag} = E(\hat{N}_{Ag})$ . Notons que le total de population peut également être exprimé sous la forme  $t_y = \sum_g t_{yg}$ .

Les composantes du biais sont nulles uniquement sous certaines conditions. *i*) Si  $N_{Bg} = N_g$  pour tout  $g$  (c'est-à-dire aucune erreur de couverture dans la base de sondage repère), le biais dépend uniquement de l'association entre la variable et les propensions à la couverture,  $\text{Cov}(y_g, \phi_{Ag})$ . La valeur de Biais( $t_{yp}$ ) se réduit alors à la formule donnée dans Kim, Li et Valliant (2007, équation 2) pour l'estimateur poststratifié classique,  $t_{yps}$ . *ii*) Si les probabilités de couverture sont constantes dans chaque poststrate (c'est-à-dire  $\phi_{Ak} = \bar{\phi}_{Ag}$  pour tout  $g$ ), la deuxième composante du biais est nulle. Uniquement si les deux conditions sont satisfaites pouvons-nous dire que  $t_{yp}$  est approximativement sans biais. D'aucuns soutiendront peut-être que l'on pourrait former une combinaison « parfaite » de poststrates telle que les composantes positives et négatives s'annulent. Cependant, nous pensons que la probabilité qu'une telle combinaison existe est tellement faible qu'elle est virtuellement impossible.

Ayant examiné le biais, nous présentons une évaluation de la variance de  $t_{yp}$ . Pour certains estimateurs, la

population connu (par exemple  $N_g^e$ ). Si l'on estime le nombre d'éléments dans la poststrate  $g$  en posant que  $y_k = 1$  dans la formule pour  $t_{Ayg}$ ,  $t_{ypS}$  égale  $N_g^e$ . En ce sens,  $t_{ypS}$  est poststratifié en fonction des chiffres de population  $N_1, \dots, N_G$ .

Cependant, dans certaines situations, les chiffres de population ne sont pas disponibles et doivent être estimés d'après une enquête repère. Exprimons l'estimateur PSCF d'un total de population d'une variable  $y$  sous la forme

$$t_{yp} = \sum_{g=1}^G \hat{N}_{Bg} \frac{t_{Ayg}}{N_{Ag}}. \quad (3)$$

Le nombre d'éléments de la population dans la  $g^e$  poststrate ( $g = 1, \dots, G$ ) estimé d'après l'enquête repère est désigné par  $\hat{N}_{Bg} = \sum_{i \in s_g} w_i$ , où  $s_{Bg}$  est l'ensemble d'éléments de l'échantillon dans la poststrate  $g$  provenant de l'enquête repère et  $w_i$  est le poids associé au  $i^e$  élément. Les facteurs de correction par calage appliqués aux poids de sondage de l'enquête analytique pour  $t_{yp}$  sont calculés selon l'expression  $a_k = \hat{N}_{Bg}/N_{Ag}$  pour  $k \in s_{Ag}$ .

Si nous relient les estimateurs poststratifiés au système de calage décrit à la section précédente,  $t_{Ax}$  est un vecteur de longueur  $G$  de chiffres de population estimés pour chaque poststrate, tel que  $t_{Ax} = (t_{Ax1}, \dots, t_{AxG})$ , où  $t_{Axg} \equiv \hat{N}_{Ag} = \sum_{k \in s_{Ag}} d_k \delta_{gk}$  et  $x_k = \delta_{gk} = 1$  si l'élément  $k$  appartient à la  $g^e$  poststrate et 0 autrement. Le vecteur  $t_{Lx}$  correspond soit à  $\mathbf{N} = (N_1, \dots, N_G)$  pour l'estimateur  $t_{yps}$  donné par (2), soit à  $\hat{\mathbf{N}}^B = (\hat{N}_{B1}, \dots, \hat{N}_{BG})$ , un vecteur de dimension  $G \times 1$  d'estimations de contrôle d'après l'enquête repère pour l'estimateur  $t_{yp}$  donné par (3).

L'estimateur  $t_{yp}$  peut être exprimé en notation matricielle sous la forme  $t_{yp} = \hat{\mathbf{N}}_A^B \hat{\mathbf{V}}_A$ , où  $\hat{\mathbf{V}}_A = (\hat{\mathbf{N}}_A)^{-1} t_{Ay}$ , un vecteur de dimension  $G \times 1$  d'estimations d'après l'enquête analytique de la forme  $\hat{\mathbf{V}}_A = [t_{A1}/\hat{N}_{A1}, \dots, t_{AG}/\hat{N}_{AG}]$ ;  $\hat{\mathbf{N}}_A = \text{diag}(\hat{N}_{A1}, \dots, \hat{N}_{AG})$ , une matrice diagonale de totaux de poststrate estimés d'après l'enquête analytique, et  $t_{Ay} = [t_{A1}, \dots, t_{AG}]$  est un vecteur de dimension  $G \times 1$  de totaux de poststrate pour la variable de résultat estimés d'après l'enquête analytique. Les autres variables associées à la notation matricielle ont été définies plus haut.

Une correction par poststratification efficace peut réduire le biais dans les estimations ponctuelles résultantes, et réduit ou accroît à peine la variance comparativement à la pondération non corrigée. Cet effet est bien connu dans le cas de la poststratification classique; aux sections suivantes, nous présentons l'évaluation comparative dans la situation où les valeurs de contrôle sont estimées.

### 3. Biais dans la PSCF pour un total de population

La poststratification classique a la réputation de réduire le biais associé à une base de sondage incomplète. Cette



par exemple, par Valliant (1993), Rust et Rao (1996), Canty et Davison (1999), Théberge (1999), Rao et Shao (1999), Yung et Rao (1996 ; 2000), et Kott (2006).

Les auteurs des articles susmentionnés émettent l'hypothèse que les totaux de contrôle, en fonction desquels sont ajustées les estimations d'échantillon auxiliaires, sont soit des valeurs réelles de population connues sans erreur, soit des valeurs tirées d'une enquête indépendante, très précise, beaucoup plus importante que l'enquête nécessitant le calage. Dans certains cas, cependant, ces totaux de contrôle sont estimés d'après des enquêtes dont les variances d'échantillonnage ne sont pas négligeables. Par exemple, on a tenté de caler des enquêtes par panel en ligne sur des valeurs tirées d'enquêtes de référence de haute qualité, distinctes, dont la portée n'est pas beaucoup plus grande que celle des enquêtes par panel proprement dites (par exemple, Krotki 2007 ; Terharian, Bremer, Smith et Thomas 2000).

De nombreux chercheurs appliquent des formules élaborées pour la poststratification classique, même si les totaux de contrôle ont été estimés. L'hypothèse tacite est que toute erreur supplémentaire (variance et biais) associée à ces valeurs de contrôle est négligeable et peut être ignorée. À l'heure actuelle, la validité de cette hypothèse ne peut pas être vérifiée et elle ne pourra l'être que lorsque l'on aura brossé un tableau complet de la poststratification en fonction des totaux de contrôle estimés.

## 2. L'estimateur poststratifié en fonction de totaux de contrôle estimés

Afin de faciliter notre discussion de l'estimateur poststratifié en fonction de totaux de contrôle estimés, nous donnons à l'enquête nécessitant la poststratification le nom d'*enquête analytique* et à la source des totaux de contrôle le nom d'*enquête repère*. En pratique, les totaux de contrôle peuvent être tirés d'une enquête repère. Cependant, pour le développement théorique, nous supposons qu'une seule de ces enquêtes est utilisée, afin qu'il soit possible d'estimer les variances et les covariances des totaux de contrôle.

Soit  $U$  la population finie cible contenant  $N$  éléments et  $t_y = \sum_{k \in U} y_k$ , le total de population d'intérêt d'une variable  $y$ . Soit  $s_A$  un échantillon aléatoire de taille  $n_A$  tiré de la base de sondage  $U_A$  pour l'enquête analytique. Un échantillon aléatoire  $s_B$  de taille  $n_B$  est sélectionné pour l'enquête repère dans la base de sondage correspondante  $U_B$ . Nous permettons qu'il soit possible que chacune des bases de sondage,  $U_A$  et  $U_B$ , ne couvre pas entièrement la population cible  $U$ . Cependant, la couverture est traitée comme un événement aléatoire, de sorte que tous les éléments compris dans la population cible ont une probabilité

positive d'être couverts par la base de sondage analytique ou par la base de sondage repère.

Dans tout l'exposé, nous adoptons la convention qu'un indice inférieur « A » signifie une association avec l'enquête analytique, telle qu'un paramètre du plan de sondage ou une estimation. Un indice inférieur « B » désigne les quantités associées à l'enquête repère. Ces indices inférieurs sont absents des paramètres associés à la population étudiée, c'est-à-dire  $t_y$ .

Sous le plan de sondage stratifié à plusieurs degrés supposé pour l'enquête analytique,  $m_{Ah} (m_{Ah} \geq 2)$  unités primaires d'échantillonnage (UPB), désignées par l'indice  $i$ , sont sélectionnées avec remise parmi un total de  $M_{Ah}$  UPB dans la  $h^{\text{e}}$  strate du plan de sondage ( $h = 1, \dots, H$  avec  $H \geq 2$ ). Nous supposons que les  $n_{Ahi}$  éléments, portant chacun l'indice  $k$ , sont tirés de  $N_{Ahi}$  dans l'UPB  $hi$  de façon qu'il soit possible de produire une estimation sans biais du total de l'UPB. Le poids de sondage,  $d_k$ , est calculé comme l'inverse de la probabilité d'inclusion inconditionnelle pour  $k \in s_{Ahi}$ , l'ensemble d'éléments de l'enquête analytique dans la  $hi^{\text{e}}$  UPB. Donc,  $n_A$ , la taille de l'échantillon de l'enquête analytique, est calculée comme  $n_A = \sum_{h=1}^H \sum_{i=1}^{m_{Ah}} n_{Ahi}$ . Dans le cas de l'enquête repère, les éléments sont tirés aléatoirement de la base de sondage correspondante ; aucune spécification explicite n'est faite pour la méthode d'échantillonnage aléatoire.

La poststratification peut être utilisée pour corriger les erreurs d'échantillonnage et de couverture. Par conséquent, nous permettons l'existence d'un sous-dénombrement dans les bases de sondage de l'enquête analytique ainsi que de l'enquête repère. En outre, nous ne tenons pas compte des effets de la non-réponse.

Supposons que la population  $U$  peut être divisée en  $g = 1, \dots, G$  poststrates mutuellement exclusives et exhaustives. Quand le nombre d'éléments dans la population,  $N^g$ , est connu pour chaque poststrate, l'estimateur poststratifié classique d'un total pour  $y$  est défini par l'expression

$$(2) \quad t_{yPS} = \sum_{g=1}^G N^g \frac{t_{yAg}}{N_{yAg}},$$

où  $y_k$  est la valeur de la variable d'analyse  $y$  pour l'élément  $k$  ;  $t_{yAg} = \sum_{k \in s_A} y_k d_k$ , le total de  $y$  dans la poststrate  $g$  estimée d'après les données de l'enquête analytique ;  $N_{yAg} = \sum_{k \in s_A} d_k$ , le total estimé d'après l'enquête analytique dans la poststrate  $g$ , et  $\delta_{yAg} = 1$  si l'élément appartient à la  $g^{\text{e}}$  poststrate et est égal à zéro autrement. Notons que  $t_{yAg}$  peut également s'exprimer sous la forme  $t_{yAg} = \sum_{k \in s_A} d_k y_k$ , où  $s_{yAg}$  désigne l'ensemble d'éléments de l'enquête analytique dans la poststrate  $g$ . Dans cette dernière expression, nous utilisons la notation « chapeau » pour faire la distinction entre un estimateur de population (par exemple  $N_{yAg}$ ) et le paramètre de

appliqué au  $k^e$  élément est défini comme une fonction du poids de sondage,  $d_k$ , et d'un facteur de correction par calage,  $a_k$ , également appelé poids  $g$  (Särndal et coll. 1992). Les poids de calage sont calculés en minimisant une fonction spécifiée qui mesure la distance entre les poids de sondage et les poids de calage sous un ensemble de contraintes définies comme étant :

$$(1) \quad \mathbf{t}^{LX} = \mathbf{t}^{Ax}$$

où  $\mathbf{t}^{LX} = \sum_{k \in U} \mathbf{x}_k$ , le vecteur de totaux de contrôle (dénom-  
brements) de population correspondant aux  $G$  ( $G \geq 1$ )  
variables auxiliaires,  $\mathbf{t}^x = \sum_{k \in s} w_k \mathbf{x}_k$ , les totaux de contrôle  
de population estimés correspondant aux composantes de  
 $\mathbf{t}^{Lx}$ , et  $\mathbf{x}_k$  est un vecteur de longueur  $G$  contenant les  
valeurs des variables auxiliaires ou d'étalonnage pour  
l'élément  $k$ . Notons que  $\mathbf{x}_k$  peut contenir des valeurs un et  
des valeurs zéro pour indiquer la présence ou l'absence  
d'une caractéristique donnée (par exemple, âge de 18 à 25  
ans), ou des valeurs plus grandes (par exemple nombre  
d'enfants). La fonction de distance des moindres carrés  
généralisés (ou du khi-deux)  $\sum_{k \in s} (w_k - d_k)^2 / c_k d_k$  qui est  
minimisée sous les contraintes données par (1) est un  
exemple d'un tel système de calage. Ce système produit une  
solution analytique appelée estimateur par la régression  
généralisée (GREG) pour  $c_k = 1$  (Deville et Särndal 1992).  
L'estimateur poststratifié est un cas particulier de l'esti-  
mateur GREG.

Les méthodes d'estimation de la variance pour l'estima-  
teur poststratifié, et de façon plus générale pour l'estimateur  
GREG, ont été étudiées abondamment. Binder (1995)  
démontre les méthodes utilisées pour calculer un estimateur  
de variance par *linéarisation de Taylor* pour l'estimateur  
GREG. D'autres références pour l'estimateur de variance  
par linéarisation sous poststratification (et sous calage de  
manière plus générale) comprennent Deville, Särndal et  
Sautory (1993), Demnati et Rao (2004), ainsi que  
Hidiroglou et Patak (2006), Särndal, Swensson et Wretman  
(1989) ont élaboré une estimation de la variance approxi-  
mative par linéarisation de l'estimateur GREG d'un total de  
population sous forme d'une fonction des résidus de popu-  
lation issus d'un modèle spécifié et des poids de sondage  
( $d_k$ ). Valliant (1993), ainsi que Yung et Rao (1996) ont  
modifié l'estimateur de variance fondé sur les résidus en  
multipliant les résidus d'échantillon par les poids de calage  
 $w_k (= a_k d_k)$ . Ils ont démontré que cet estimateur révisé,  
créé en linéarisant l'estimateur jackknife connexe, réduisait  
le biais associé à la formule originale. Cet estimateur de  
variance est également discuté dans Särndal et coll. (1992),  
Stukel, Hidiroglou et Särndal (1996), ainsi que dans le  
chapitre 11 de Särndal et Lundström (2005). Les propriétés  
des estimateurs de variance par répliques (c'est-à-dire jack-  
knife et répliques répétées équilibrées) ont été examinées,

sur des estimations démographiques provenant de la CPS de  
mars 2007, ainsi que sur des estimations des habitudes d'uti-  
lisation du téléphone établies d'après la *National Health  
Interview Survey* réalisée de juillet à décembre 2007  
(Keeter, Dimock et Christian 2008).

Le but de notre étude est d'élaborer et d'évaluer des  
estimateurs de variance pour des estimations ponctuelles  
calculées en se servant de poids qui contiennent une cor-  
rection par poststratification en fonction d'un ensemble de  
totaux de contrôle estimés par sondage. Nous donnons à la  
méthodologie qui tient compte correctement des totaux de  
contrôle estimés le nom de *poststratification en fonction de  
totaux de contrôle estimés* (CE). Dans le présent article, nous  
nous intéressons tout spécialement à l'estimateur poststratifié  
en fonction de totaux de contrôle estimés (PSCF) d'un total  
de population pour des données recueillies selon un plan de  
sondage stratifié à plusieurs degrés, où les unités d'échan-  
tillonnage de premier degré sont sélectionnées *avec remise*.  
Dans la suite de la présente section, nous passons brièvement  
en revue le calage des poids et la poststratification. La  
section 2 contient une définition explicite de l'estimateur  
PSCF étudié et la section 3 donne une évaluation de ses  
propriétés de biais. Au moyen d'une évaluation théorique  
(section 4) et d'une étude par simulation, nous comparons les  
estimateurs de variance élaborés pour l'estimateur PSCF à  
un estimateur de variance choisi sous l'hypothèse naïve des  
« totaux de contrôle de population ». Notre étude porte à la  
fois sur les estimateurs de variance par linéarisation et par  
répliques. Nous donnons des exemples des effets qu'ont sur  
les estimations de la variance divers niveaux de précision des  
totaux de contrôle estimés. À la section 5, nous décrivons en  
détail les spécifications de l'étude par simulation et à la  
section 6, nous résumons les résultats. L'article se termine  
par un bref résumé et un aperçu des futurs travaux de  
recherche dans le domaine

Les *estimateurs par calage* (Deville et Särndal 1992),  
tels qu'un estimateur poststratifié d'un total de population,  
empruntent de l'information auxiliaire pour accroître l'effi-  
cacité des estimations par sondage comparativement aux  
méthodes de pondération plus simples. Si les variables  
auxiliaires sont reliées (linéairement) à l'ensemble de varia-  
bles étudiées clés, les estimateurs par calage peuvent être  
très efficaces.

La forme générale d'un estimateur par calage *classique*  
ou *sur totaux de contrôle fixes* se décrit le mieux comme un  
estimateur à facteur d'extension ou « à pondération liné-  
aire » tel qu'il est décrit dans Estevao et Särndal (2000).  
Designons par  $s$  l'ensemble d'éléments d'échantillon pro-  
venant d'un échantillon probabiliste et par  $d_k = 1/\pi_k$  le  
poids de sondage de l'élément  $k$  tel que  $\pi_k = \Pr(k \in s)$ .  
Un total de population estimé d'une variable  $y$  est donné  
par  $t_y^v = \sum_{k \in s} w_k y_k$ , où le poids de calage ( $w_k = a_k d_k$ )



# Une comparaison des estimateurs de la variance pour la poststratification en fonction de totaux de contrôle estimés

Jill A. Dever et Richard Valliant<sup>1</sup>

## Résumé

Les méthodes de calage, telles que la poststratification, s'appuient sur de l'information auxiliaire pour accroître l'efficacité des estimations par sondage. L'hypothèse est que les totaux de contrôle, en fonction desquels les poids de sondage sont poststratifiés (ou calés), sont les valeurs de population. Toutefois, les totaux de contrôle sont souvent estimés d'après d'autres enquêtes. De nombreux chercheurs appliquent les estimateurs classiques d'estimation de la variance par poststratification à des situations où les totaux de contrôle sont estimés, supposant donc que toute variance d'échantillonnage supplémentaire associée à ces totaux estimés est négligeable. Le but de l'étude présentée ici est d'évaluer des estimateurs de la variance pour des plans de sondage stratifiés à plusieurs degrés, sous une poststratification en fonction de totaux de contrôle estimés (CE) en utilisant des valeurs de contrôle sans biais par rapport au plan. Nous comparons les propriétés théoriques et empiriques des estimateurs de variance par linéarisation et par le jackknife pour un estimateur poststratifié d'un total de population. Nous donnons des exemples des effets qu'ont sur les variances divers niveaux de précision des totaux de contrôle estimés. Notre étude donne à penser que i) les estimateurs de variance classiques peuvent sous-estimer considérablement la variance théorique et que ii) deux estimateurs de variance par poststratification CE peuvent atténuer le biais négatif.

Mots clés : Poststratification en fonction de totaux de contrôle estimés ; biais de couverture de la base de sondage ; totaux de contrôle estimés d'après un sondage.

## 1. Introduction

Les estimateurs poststratifiés, et d'autres estimateurs par calage, sont utilisés dans de nombreux types de sondage pour réduire les variances ou pour corriger certains défauts de la base de sondage. À titre d'exemple particulier, mentionnons les grandes enquêtes du gouvernement des États-Unis, telles que la Consumer Expenditure Survey (voir, par exemple, Jayasuriya et Valliant 1996), les enquêtes auprès de populations spéciales, telles que la Survey of Health Related Behaviors among Military Personnel du U.S. Department of Defense (Bray, Hourani, Rae, Dever, Brown, Vincus, Pemberton, Marsden, Faulkner et Vandermas-Peeler 2003), et une foule d'enquêtes réalisées en dehors des États-Unis, dont l'Enquête sur le commerce de détail du Canada (voir, par exemple, Hidiroglou et Patak 2006), l'Enquête sur la population active de la Suède (Mitza et Hömgen 2002), et la Household Panel Survey du Royaume-Uni (Taylor, Brice, Buck et Prentice-Lane 2007).

Les estimateurs par calage, tels ceux produits par poststratification, sont utilisés pour réduire au minimum les erreurs associées aux bases de sondage incomplètes (par exemple, le sous-dénombrement) ainsi qu'à l'échantillonnage et à la non-réponse (voir, par exemple, Särndal, Swensson et Wretman 1992; Lessler et Kalsbeek 1992; Kott 2006). Ainsi, les estimations produites d'après le Behavioral Risk Factor Surveillance System (BRFSS), une enquête téléphonique par composition aléatoire (CA) de

S'il n'existe pas de totaux de contrôle de population pertinents, de nombreux chercheurs utilisent des totaux de contrôle estimés d'après un sondage et appliquent les formules classiques de calcul de la variance comme si les totaux de contrôle étaient connus sans erreur. Par exemple, Nadimpalli, Judkins et Chu (2004) ont corrigé les poids pour la *National Survey of Parents and Youth de 2003* en fonction du nombre de ménages américains comptant des enfants de 9 à 18 ans estimé d'après la *Current Population Survey (CPS)* en se servant d'un algorithme de ratisage (*raking ratio*) ([www.census.gov/cps](http://www.census.gov/cps)). Des estimations de la façon dont les personnes vivant aux États-Unis emploient leur temps peuvent être calculées d'après l'*American Time Use Survey* en utilisant des poids qui ont été poststratifiés en fonction d'estimations projetées d'après le recensement décennal des États-Unis (Killion 2006). Plus récemment, aux Pew Research Centers, des chercheurs ont calé des poids pour un ensemble de sondages pré-électoraux réalisés durant la campagne présidentielle de 2008 aux États-Unis

et Valliant 2007).

portée nationale réalisée par les Centers for Disease Control and Prevention (CDC) des États-Unis, sont poststratifiées en fonction des nombres de ménages équipés et non équipés d'un service téléphonique classique à fil (Centers for Disease Control and Prevention 2006). La réduction des erreurs est liée à l'association des totaux de contrôle de population avec le sous-dénombrement de la base de sondage, les profils de non-réponse non-ignorable et la variable d'intérêt (Kim, Li

1. Jill A. Dever, RTI International, Courtiel : jdever@rti.org ; Richard Valliant, Survey Research Center, University of Michigan and Joint Program in Survey Methodology, University of Maryland, Courtiel : rvalliant@survey.umd.edu.



Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics*, 20, 1-18.

Yung, W., et Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95, 903-915.

- Brick, M.J., et Montaquila, J.M. (2009). Nonresponse and weighting. Dans *le Handbook of Statistics, Sample Surveys: Design, Methods and Applications*, (Eds., C.R. Rao et D. Pfeffermann), 29A, 163-185.
- Da Silva, D.N., et Opsomer, J.D. (2006). A kernel smoothing method to adjust for unit nonresponse in sample surveys. *Canadian Journal of Statistics*, 34, 563-579.
- Deville, J.-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des Journées de Méthodologie Statistique*, Insee.
- Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Ehinge, J.L., et Vansaneh, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23, 37-45.
- Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429-440.
- Katton, G., et Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.
- Kim, J.K., et Kim, J.I. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501-514.
- Kott, P. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 149-160.
- Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.
- Ozcoskun, L., Thompson, K.J., et Williams, Q. (2005). Investigation of balanced repeated replication (BRR) variance estimation for the Survey of Residential Alterations and Repairs (SORAR). *Proceedings of the Federal Committee on Statistical Methodology*, Office of Management and Budget.
- Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York : John Wiley & Sons, Inc.
- Sastry, O. (2003). CALMAR 2 : une nouvelle version du programme CALMAR de redressement d'échantillon par calage. *Recueil : Symposium 2003, Défis reliés à la réalisation d'enquêtes pour la prochaine décennie*, Ottawa, Canada.
- Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 93, 254-265.
- Thompson, K.J. (2005). An empirical investigation into the effects of replicate reweighting on variance estimates for the annual capital expenditures survey. *Proceedings of the Federal Committee on Statistical Methodology*, Office of Management and Budget.
- Thompson, K.J., et Yung, W. (2006). To Replicate (A weight adjustment procedure) or not to replicate? An analysis of the stratified jackknife. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3772-3779.
- Statistique Canada, N° 12-001-X au catalogue
- Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society B*, 67, 445-458.
- Bibliographie**
- Le présent rapport est publié en vue d'informer les parties intéressées des travaux de recherche en cours et de favoriser la discussion de ces travaux. Toutes opinions exprimées concernant les problèmes statistiques, méthodologiques ou opérationnels sont celles des auteurs et ne reflètent pas forcément celles du U.S. Census Bureau. Les auteurs remercient le rédacteur associé, deux examinateurs anonymes, Samson Adeshiyun, Patrick Cantwell, Carol Caldwell, Michael Hidiroglou, Rita Petroni, Mark Sands et Jun Shao de leurs commentaires constructifs concernant des versions antérieures du présent article. Les travaux de David Haziza ont été financés par des bourses du Conseil de recherches en sciences naturelles et en génie du Canada.
- Remerciements**
- L'une des justifications de l'utilisation d'une procédure simplifiée dans une méthode d'estimation de la variance par rééchantillonnage consiste à gagner du temps et à économiser les ressources informatiques. S'il s'agit de considérations vraiment importantes et que le programme obtient systématiquement des taux de réponse élevés pour les unités dans toutes les cellules de pondération, alors que la production de répliques de la procédure d'ajustement des poids présente des avantages théoriques manifestes, les avantages pratiques pourraient être peu nombreux, voir inexistants. Cela dit, les conditions d'équivalence « pratique » entre les estimateurs de variance par les méthodes complète et simplifiée sont extrêmement contraignantes et nous avons démontré que de faibles variations des conditions relatives aux données sous-jacentes peuvent facilement donner lieu à une violation de ces conditions d'équivalence. Si le jackknife pose réellement un problème en ce qui concerne les ressources informatiques, les auteurs recommandent l'approche de l'estimation jackknife par linéarisation de la variance, qui a les mêmes propriétés asymptotiques que le jackknife complet, mais dont les calculs sont rapides et dont le temps inactif du système est « gratuit » (en ce qui concerne la mise en mémoire des répliques). Voir Thompson et Yung (2006) pour des expressions de l'estimateur de variance jackknife par linéarisation pour les estimateurs ajustés par la fréquence ainsi que par le ratio. Etant donné ces alternatives viables, nous déconseillons d'utiliser un estimateur de variance à procédure simplifiée.

et 90 % dans la strate 3. Ce profil de réponse n'est pas inhabituel dans les enquêtes-entreprises où un suivi plus important est effectué pour les unités de moyenne et de grande taille (strates 2 et 3).

Pour chaque échantillon, nous avons calculé les estimateurs ajustés par la fréquence et par le ratio, donnés respectivement par (1.4) et (1.5), en utilisant les strates comme classes de pondération. Nous avons estimé la variance des estimateurs ponctuels au moyen de  $v_{JP}$  et de  $v_{JS}$ , donnés respectivement par (2.2) et (2.3). Comme mesure du biais d'un estimateur de variance  $v$ , nous avons utilisé le biais relatif Monte Carlo en pourcentage donné par

$$BR_{MC}(v) = \frac{1}{5\,000} \sum_{i=1}^{5\,000} \frac{E\hat{Q}M_{MC}(Y_{CAL}^{(i)}) - E\hat{Q}M_{MC}(Y_{CAL})}{E\hat{Q}M_{MC}(Y_{CAL})} \times 100.$$

où  $v^{(i)}$  est l'estimation de la variance obtenue à partir du  $i^e$  échantillon, et  $E\hat{Q}M_{MC}(Y_{CAL})$  est l'erreur quadratique moyenne ( $E\hat{Q}M$ ) Monte Carlo définie par

$$E\hat{Q}M_{MC}(Y_{CAL}) = \frac{1}{50\,000} \sum_{i=1}^{50\,000} (Y_{CAL}^{(i)} - Y)^2,$$

où  $Y_{CAL}^{(i)}$  est l'estimation (corrigée par le ratio ou par la fréquence) de  $Y$  pour le  $i^e$  échantillon. Le tableau 3 donne le biais relatif Monte Carlo en pourcentage pour les estimateurs ajustés par la fréquence ainsi que par le ratio.

**Tableau 3**  
Biais relatif Monte Carlo en pourcentage pour les estimateurs de variance jackknife simplifié et complet

Population	Estimateur ajusté par la fréquence	Estimateur ajusté par le ratio
1	57,3 % $BR_{MC}(v_{JS})$	1,1 % $BR_{MC}(v_{JP})$
2	877,1 % 220,7 %	364,7 % 185,9 %
3	21,6 % 266,4 %	0,6 % 0,2 %
4	29,1 % -67,2 %	1,4 % 5,0 %
5		

Comme prévu, l'estimateur simplifié surestime l'EQM Monte Carlo de l'estimateur ajusté par la fréquence pour toutes les populations. La surestimation varie d'environ 20 % pour la population 4 à plus de 800 % pour la population 2. L'expression (3.8) montre que le biais de  $v_{JS}$  dépend du taux de réponse et de  $\bar{y}_h^2$ . Pour la population 2, le terme d'ordonnée à l'origine est grand et accroît la valeur de  $CV_h(v)$  dans toutes les strates, ce qui à son tour accroît le biais de  $v_{JS}$ . La population 3 est similaire à la population 2, excepté que le terme d'ordonnée à l'origine n'est grand que pour la première strate. Comme prévu, le biais de  $v_{JS}$  dans cette population est compris entre ceux des populations 1 et 2. La population 4 est celle qui imite la population de l'ACES avec les valeurs remplacées par zéro pour certaines unités dans les strates 1 et 2. Le biais relatif Monte Carlo de

21,6 % émane, en grande partie, de la troisième strate où aucune unité n'a été remplacée par une valeur nulle (ce fait peut être constaté en utilisant l'expression (3.8)). Par comparaison, pour chacune des cinq populations, l'estimateur de variance jackknife complet suit très bien l'EQM Monte Carlo, le biais relatif absolu étant inférieur à 1,1 %. Si nous examinons l'estimateur ajusté par le ratio, nous voyons que, de nouveau, l'estimateur de variance jackknife suit relativement bien l'EQM Monte Carlo pour toutes les populations, le biais relatif absolu étant inférieur à 5 %. Par contre, le biais relatif de l'estimateur simplifié varie de -67 % à 364 %. L'examen de l'expression (3.7) montre que, pour un taux de réponse fixe, le biais dépend de  $CV_h(v)$ ,  $CV_h(x)$  et  $p_{hxy}$ . Étant donné l'importance des termes d'ordonnée à l'origine dans la deuxième population, les valeurs de  $\bar{y}_h$  sont grandes et les  $CV_h(v)$  correspondants sont plus faibles que pour les autres populations. Donc, le dernier terme de l'expression (3.7) est assez grand et le biais relatif résultant de  $v_{JS}$  est également grand. Nous faisons la même constatation pour la population 3, mais à un degré moindre, puisque seule la première strate possède un terme d'ordonnée à l'origine. Nous observons l'effet opposé dans la population 4, où l'introduction de valeurs nulles accroît significativement  $CV_h(v)$ , ce qui à son tour réduit le biais relatif Monte Carlo en pourcentage de l'estimateur simplifié.

Des simulations supplémentaires ont été exécutées en utilisant certaines populations décrites au tableau 2, mais en faisant varier les taux de réponse. Les résultats ne sont pas présentés ici car ils correspondaient à ceux attendus. Autrement dit, le biais de l'estimateur simplifié diminuait à mesure que le taux de réponse augmentait (tous les autres paramètres étant maintenus constants). L'estimateur jackknife complet continuait de très bien suivre l'EQM Monte Carlo.

## 5. Conclusion

Dans le présent article, nous avons évalué théoriquement et empiriquement un estimateur de variance jackknife simplifié dans lequel les facteurs d'ajustement pour la non-réponse ne sont pas recalculés dans chaque réplique jackknife, en étudiant en particulier trois procédures distinctes d'ajustement de la pondération pour la non-réponse. Nous avons montré, dans le contexte de l'échantillonnage aléatoire simple stratifié, que l'estimateur de variance jackknife simplifié a tendance à surestimer la variance réelle des estimateurs. Toutefois, dans le contexte de la procédure d'ajustement par le ratio, cet estimateur pourrait sous-estimer la variance réelle si le modèle du ratio n'est pas adapté aux données dont on dispose.



Dans le cas de l'estimateur ajusté par le ratio, l'expression (3.2) peut être étendue facilement au cas de l'échantillonnage aléatoire simple stratifié pour obtenir

$$\text{Biais}(v_{JS}) \approx E_{pq}(\tilde{D}) \approx \sum_{h=1}^L \frac{N_h^2}{N^2} E_{pq} \left( r_h \right) \left( 1 - E_{pq} \left( \frac{r_h}{r_h} \right) \right) \left( S_y^2 \frac{CV_h(y)^2}{1} + 2 \frac{CV_h(x)}{CV_h(y)} p_{hxy} - \frac{CV_h(x)^2}{CV_h(y)^2} \right), \quad (3.7)$$

où les quantités  $r_h$ ,  $CV_h(x)$ ,  $CV_h(y)$ ,  $S_y^2$  et  $p_{hxy}$  correspondent à  $r$ ,  $CV(x)$ ,  $CV(y)$ ,  $S_y^2$  et  $p_{hxy}$  calculés dans chaque strate.

Pour l'estimateur ajusté par la fréquence, l'expression (3.4) peut être étendue facilement au cas de l'échantillonnage aléatoire simple stratifié pour obtenir

$$\text{Biais}(v_{JS}) \approx E_{pq}(\tilde{D}) \approx \sum_{h=1}^L \frac{N_h^2}{N^2} E_{pq} \left( r_h \right) \left( 1 - E_{pq} \left( \frac{r_h}{r_h} \right) \right) \left( S_y^2 \frac{CV_h(y)^2}{1} + 2 \frac{CV_h(x)}{CV_h(y)} p_{hxy} - \frac{CV_h(x)^2}{CV_h(y)^2} \right), \quad (3.8)$$

Enfin, pour l'estimateur ajusté par régression linéaire simple, l'expression (3.6) peut être étendue facilement au cas de l'échantillonnage aléatoire simple stratifié pour obtenir

$$\text{Biais}(v_{JS}) \approx E_{pq}(\tilde{D}) \approx \sum_{h=1}^L \frac{N_h^2}{N^2} E_{pq} \left( r_h \right) \left( 1 - E_{pq} \left( \frac{r_h}{r_h} \right) \right) \left( S_y^2 \frac{CV_h(y)^2}{1} + 2 \frac{CV_h(x)}{CV_h(y)} p_{hxy} - \frac{CV_h(x)^2}{CV_h(y)^2} \right), \quad (3.9)$$

Des expressions (3.7) à (3.9), il découle qu'il faut faire preuve d'une certaine prudence lorsque l'on utilise l'estimateur de variance jackknife simplifié. En effet, même si le biais de cet estimateur est faible dans chaque strate, la somme de ces biais au niveau de la population peut être considérable s'ils ont tous la même direction.

#### 4. Étude par simulation

Une étude par simulation nous a permis de comparer les propriétés statistiques des estimateurs de variance jackknife simplifiée et complet sous diverses conditions. Nous avons généré cinq populations stratifiées différentes contenant chacune 30 000 unités et deux variables. D'abord, nous avons tiré les valeurs de  $x$  d'une loi gamma dont les paramètres étaient  $\alpha$  et  $\lambda$ . Puis, sachant les valeurs de  $x$ , nous avons généré les valeurs de  $y$  selon le modèle suivant :

$$y_{hi} = \beta_0 + \beta_1 x_{hi} + \varepsilon_{hi},$$

où  $\varepsilon_{hi} \sim N(0, \sigma_{\varepsilon}^2)$ . Nous avons fixé les valeurs de la variance et de  $\sigma_{\varepsilon}^2$  de telle façon que le coefficient de corrélation (désigné par  $\rho_{xy}$ ) entre  $x_{hi}$  et  $y_{hi}$  soit égal à 0,7 dans toutes les populations. Chaque population a été stratifiée en trois strates contenant chacune 10 000 unités. Les paramètres des populations simulées sont présentés au tableau 2.

La population 1 correspond très bien au modèle du ratio avec une ordonnée à l'origine nulle dans toutes les strates. La population 2 possède un terme d'ordonnée à l'origine non négligeable dans chacune des trois strates. La population 3 est une combinaison des populations 1 et 2, où le modèle du ratio est bien ajusté pour les strates 2 et 3, mais non pour la strate 1. La population 4 est semblable à la population 1, excepté que les unités des strates 1 et 2 ont 70 % de chance de déclarer une valeur nulle. Cette population est destinée à imiter la situation de l'Annual Capital Expenditures Survey (ACES) du U.S. Census Bureau, qui a motivé la présente étude. L'ACES emploie un estimateur de variance jackknife simplifié qui, selon les études empiriques, donne des résultats qui s'approchent de ceux de l'estimateur de variance jackknife complet. Sa population est caractérisée par de nombreuses valeurs nulles déclarées pour les dépenses en immobilisations par la majorité des petites et moyennes entreprises échantillonnées, la majorité des dépenses déclarées étant fournie par les grandes entreprises. Nous avons généré la population 5 afin de montrer que, pour l'estimateur ajusté par le ratio, l'estimateur simplifié peut effectivement présenter un biais négatif quand le modèle du ratio est spécifié incorrectement (démontré dans l'expression (3.3) pour un échantillonnage aléatoire simple). Pour cette population, le terme d'ordonnée à l'origine est fortement significatif dans toutes les strates.

Tableau 2  
Paramètres de population

Population	$\beta_0$	$\beta_1$	$\alpha$	$\lambda$	$CV(x)$	$CV(y)$
(À l'intérieur de la strate)						
1						
1	0	2	4	5	50 %	76 %
2	120	2	4	5	50 %	44 %
3	120	2	4	5	50 %	51 %
4	0	2	4	5	50 %	134 %
5	50	0,5	1	2	200 %	63 %
2						
1						
1	0	2	4	5	50 %	76 %
2	240	2	4	5	50 %	44 %
3	360	2	4	5	50 %	51 %
4	0	2	4	5	50 %	134 %
5	200	0,5	1	2	200 %	63 %

De chaque population, nous avons tiré 5 000 échantillons aléatoires simples stratifiés de taille 300 (100 unités par strate). Dans chaque échantillon, nous avons produit une non-réponse en utilisant un mécanisme de réponse uniforme à l'intérieur de chaque strate, avec les probabilités de réponse égales à 60 % dans la strate 1, 70 % dans la strate 2

à condition que  $0 < E_{pq}(r/n) < 1$ , où  $B_0 = \bar{Y} - B_1\bar{X}$  est l'ordonnée à l'origine en population finie de la droite des moindres carrés lorsque l'on fait la régression de  $y$  sur  $x$

avec

$$B_1 = \frac{\sum_{i \in U} (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i \in U} (x_i - \bar{X})^2}.$$

L'expression (3.2) montre clairement que le biais de  $v_{JS}$

augmente si i) le taux de réponse prévu  $E_{pq}(r/n)$  diminue, ii)  $p_{xy}$  augmente, iii)  $CV(y)$  diminue ou iv)  $CV(x)$  augmente. En outre, il découle de (3.3) que  $v_{JS}$  surestime la variance réelle quand l'ordonnée à l'origine  $B_0$  n'est pas trop grande. Le tableau 1 illustre la relation entre  $CV(x)$  et la condition (3.3). Par exemple, quand  $CV(x) = 0$ ,  $v_{JS}$  surestime toujours la variance réelle puisque, dans ce cas, la condition (3.3) se réduit à la condition  $B_0 < \infty$ , qui est toujours satisfaite. Ce résultat n'est pas étonnant, car si  $CV(x) = 0$ , les valeurs de  $x$  sont toutes égales et l'estimateur ajusté par le ratio (1.5) est identique à l'estimateur ajusté par la fréquence (1.4). Comme nous le discutons plus loin,  $v_{JS}$  surestime toujours la variance réelle dans ce cas.

Quand  $CV(x)$  est grand (par exemple  $CV(x) = 2$ ),  $v_{JS}$  surestime la variance réelle si, et uniquement si,  $B_0 < 0,625\bar{Y}$ . Cette dernière condition est satisfaite si l'ordonnée à l'origine n'est pas « trop loin » de l'origine. Par conséquent, si la relation entre  $y$  et  $x$  passe par l'origine (c'est-à-dire si le modèle du ratio est vérifié), l'estimateur de variance simplifié surestime la variance réelle. Par contre, si l'estimateur ajusté par le ratio est utilisé quand le modèle du ratio n'est pas vérifié, par exemple quand  $B_0 \geq 0,625\bar{Y}$ , l'estimateur de variance simplifié  $v_{JS}$  sous-estime la variance réelle. En conclusion, nous pouvons nous attendre à ce que  $v_{JS}$  surestime la variance réelle quand nous utilisons une méthode d'ajustement par le ratio, à moins que le modèle du ratio soit extrêmement mal spécifié pour les données disponibles, ce qui pourrait se produire, par exemple, si les variables  $y$  et  $x$  sont négativement corrélées.

Tableau 1  
Relation entre  $CV(x)$  et la condition (3.3)

$CV(x)$	$\frac{\bar{Y}}{2} \left( \frac{1 + CV(x)^2}{CV(x)^2} \right)$
0	$\infty$
0,1	50,5 $\bar{Y}$
0,5	2,5 $\bar{Y}$
1	2 $\bar{Y}$
1,5	0,722 $\bar{Y}$
2	0,625 $\bar{Y}$

Si nous considérons maintenant l'estimateur ajusté par la fréquence (1.4), en posant  $x_i = 1$  pour tout  $i$  dans (3.1), nous obtenons

$$\tilde{D} = \frac{N^2}{n^2} \left( 1 - \frac{r}{r^2} \right) \frac{r}{\bar{y}^2}. \quad (3.4)$$

Il découle de (3.4) que le biais relatif de  $v_{JS}$ ,  $BR(v_{JS}) = \tilde{D}/\tilde{V}_{JP}$ , peut être approximé par  $E_{pq}(RD)$  où  $RD = \tilde{D}/\tilde{V}_{JP}$ . Sous un mécanisme de non-réponse uniforme, de simples opérations algébriques mènent à

$$BR(v_{JS}) \approx E_{pq}(RD) \approx \left( 1 - E_{pq} \left( \frac{n}{r} \right) \right) \frac{1}{CV(y)^2}. \quad (3.5)$$

L'expression (3.5) montre que, dans le cas de l'estimateur ajusté par la fréquence (1.4),  $v_{JS}$  surestime toujours la variance réelle. L'ampleur de la surestimation augmente à mesure que le taux de réponse prévu  $E_{pq}(r/n)$  diminue ou que  $CV(y)$  diminue. Par exemple, si le taux de réponse prévu est égal à 70 % et que  $CV(y) = 1$ , nous avons  $E_{pq}(RD) = 1,3$ , de sorte que la valeur de l'estimateur de variance jackknife simplifié,  $v_{JS}$ , est en moyenne 30 % plus élevée que la variance réelle de  $\tilde{Y}^{CAL}$ . Par ailleurs, si le taux de réponse est égal à 70 % et que  $CV(y) = 0,5$ , nous avons  $E_{pq}(RD) = 5,3$ , auquel cas la surestimation est considérable.

Enfin, penchons-nous sur le cas de l'estimateur ajusté par régression linéaire simple (1.6). Sous un mécanisme de non-réponse uniforme, nous pouvons montrer que le biais asymptotique de  $v_{JS}$  est donné par

$$\begin{aligned} \text{Biais}(v_{JS}) &\approx E_{pq}(D) \\ &\approx \frac{N^2}{n^2} \left( 1 - E_{pq} \left( \frac{n}{r} \right) \right) \left( S_y^2 \left( \frac{1}{CV(y)^2} + p_{xy}^2 \right) \right) \geq 0. \end{aligned} \quad (3.6)$$

De (3.6), il découle que  $v_{JS}$  surestime toujours la variance réelle dans le cas de l'estimateur ajusté par régression linéaire simple (1.6). Le biais (3.6) augmente si i) le taux de réponse prévu diminue, ii)  $p_{xy}^2$  augmente ou iii)  $CV(y)$  diminue.

### 3.2 Échantillonnage aléatoire simple stratifié : les classes de pondération coïncident avec les strates

À la présente section, nous supposons que les classes de pondération coïncident avec les strates originales du plan de sondage. Cette situation n'est pas inhabituelle en pratique, surtout dans les enquêtes-entreprises. Si les strates sont telles que les unités qu'elles contiennent ont à peu près la même propension à répondre (c'est-à-dire réponse uniforme à l'intérieur de la strate), les expressions pour le biais de  $v_{JS}$  s'obtiennent facilement à partir des expressions (3.2), (3.4) et (3.6).

### 3. Biases de $v_{JS}$ dans certains cas particuliers

#### 3.1 Échantillonnage aléatoire simple sans remise

À la présente section, nous supposons que l'échantillon  $s$  a été sélectionné selon un plan d'échantillonnage aléatoire simple sans remise. Nous supposons également que la fraction d'échantillonnage  $n/N$  est négligeable et que le nombre de répondants  $r$  est grand. Enfin, nous supposons que  $n$  existe qu'une seule classe de pondération. Bien que la situation susmentionnée ne soit pas réaliste en pratique, elle donne une certaine idée du biais asymptotique de  $v_{JS}$ . Dans le cas de l'estimateur ajusté par le ratio (1.5), nous pouvons montrer que  $D$  est donné approximativement par

$$D = \frac{N^2}{n^2} \left( 1 - \frac{n}{r} \right) \left\{ \left( \frac{\bar{x}}{s_x^2} \right) (s_x^2 - s_v^2) + 2 \left( \frac{\bar{x}}{s_x} \right) R_r \left[ \left( \frac{\bar{x}}{s_x} \right) - 1 \right] \frac{n}{s_{cv}^2} \right. \\ \left. + R_r^2 \left[ \left( \frac{\bar{x}}{s_x} \right) (s_x^2 - s_v^2) + \left( \frac{\bar{x}}{s_x} \right) \right] \right\} \quad (3.1)$$

où  $(\bar{x}_r, \bar{y}_r) = 1/r \sum_{i \in s} r_i(x_i, y_i)$  désigne la moyenne des répondants pour les variables  $x$  et  $y$  respectivement,  $R_r = \bar{y}_r / \bar{x}_r$ ,  $s_x^2 = 1/(r-1) \sum_{i \in s} r_i^2 (x_i - \bar{x}_r)^2$  avec  $\bar{x}_r = 1/n \sum_{i \in s} x_i$ ,  $s_v^2 = 1/(r-1) \sum_{i \in s} r_i^2 (y_i - \bar{y}_r)^2$  et  $s_{cv}^2 = 1/(r-1) \sum_{i \in s} r_i^2 (y_i - \bar{y}_r)^2 x_i$ . Si nous supposons en outre que toutes les unités ont la même probabilité de réponse (c'est-à-dire un mécanisme de réponse uniforme), nous avons  $\bar{x}/\bar{x}_r \xrightarrow{p} 1$  et  $s_x^2/s_{cv}^2 \xrightarrow{p} 1$ . Dans ce cas, le biais asymptotique de  $v_{JS}$  est donné par

$$\text{Biais}(v_{JS}) \approx E^{pq}(D)$$

$$\approx \frac{N^2}{r} \left( 1 - E^{pq} \left( \frac{n}{r} \right) \right)$$

$$S_v^2 \left( \frac{1}{\text{CV}(x)^2} + 2 \frac{\text{CV}(y)}{\text{CV}(x)} \rho_{xy} - \frac{\text{CV}(y)^2}{\text{CV}(x)^2} \right). \quad (3.2)$$

où  $\text{CV}(x) = S_x/\bar{X}$  et  $\text{CV}(y) = S_y/\bar{Y}$  désignent les coefficients de variation de population pour les variables  $x$  et  $y$ , respectivement avec  $S_x^2 = 1/(N-1) \sum_{i \in U} y_i^2$  et  $\bar{Y} = 1/N \sum_{i \in U} y_i$ ,  $S_y^2$  et  $\bar{X}$  sont définis de la même façon, et  $\rho_{xy}$  désigne le coefficient de corrélation en population finie pour les variables  $x$  et  $y$ . De (3.2) il découle que le biais asymptotique de  $v_{JS}$  est non négatif si, et uniquement si

$$B_0 > \frac{\bar{Y}}{2} \left( \frac{1 + \text{CV}(x)^2}{\text{CV}(x)^2} \right) \quad (3.3)$$

nous avons  $E^{pq}(v_{JS}) \neq V(Y^{\text{CAL}})$ , et l'estimateur de variance simplifié,  $v_{JS}$ , est biaisé.

Pour étudier l'ampleur du biais de  $v_{JS}$ , nous considérons la différence entre les deux estimateurs jackknife de la variance,  $D = v_{JS} - v_{JF}$ . Puisque l'estimateur de variance  $v_{JF}$  est un estimateur asymptotiquement sans biais du terme  $V(Y^{\text{CAL}} | \mathbf{r})$ , il est asymptotiquement équivalent à un estimateur de variance obtenu en utilisant un développement en série de Taylor de premier degré. L'estimateur de variance résultant, désigné par  $\tilde{v}_{JF}$ , est l'estimateur de variance jackknife par linéarisation étudié par Yung et Rao (2000). De même, l'estimateur de variance jackknife simplifié  $v_{JS}$  est asymptotiquement équivalent à un estimateur de  $V(Y^{\text{CAL}} | \mathbf{r})$  obtenu en traitant les facteurs d'ajustement des poids pour la non-réponse  $g_{hi}$  comme des constantes. Nous désignons cet estimateur de variance par  $\tilde{v}_{JS}$ . La quantité  $D$  peut donc être approximée par  $D = \tilde{v}_{JS} - \tilde{v}_{JF}$ . Pour que cette approximation soit valide, nous supposons que le nombre de répondants est grand.

En notant que  $\text{Biais}(v_{JF}) = E^{pq}(v_{JF}) - V(Y^{\text{CAL}}) \approx 0$ , il s'ensuit que le biais de  $v_{JS}$ ,  $\text{Biais}(v_{JS}) = E^{pq}(v_{JS}) - V(Y^{\text{CAL}})$ , peut être approximé par  $E^{pq}(D) \approx E^{pq}(D)$ . Soit  $v(y)$  l'estimateur de variance de l'estimateur sur données complètes (1.1). En utilisant un développement en série de Taylor de premier degré, nous pouvons montrer qu'un estimateur de  $V(Y^{\text{CAL}} | \mathbf{r})$  est donné par

$$\tilde{v}_{JF} = v(\xi) \quad (2.4)$$

où

$$\xi_{hi} = \mathbf{x}_{hi}' \mathbf{B}_r + g_{hi} r_{hi} e_{hi}$$

avec  $e_{hi} = (y_{hi} - \mathbf{x}_{hi}' \mathbf{B}_r) = \mathbf{B}_r - \mathbf{T}^{-1} \sum_{(hi) \in s} d_{hi} r_{hi} \mathbf{z}_{hi} y_{hi}$ . Par ailleurs, traiter les facteurs  $g_{hi}$  comme des constantes implique que  $Y^{\text{CAL}}$  est linéaire en les poids de sondage  $d_{hi}$ . Il s'ensuit que  $\tilde{v}_{JS}$  est donné par

$$\tilde{v}_{JS} = v(\psi), \quad (2.5)$$

où  $\psi_{hi} = g_{hi} r_{hi} y_{hi}$ . Par exemple, pour un plan d'échantillonnage à taille fixe ou aléatoire, un estimateur de variance possible est

$$\tilde{v}_{JF} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \xi_i \xi_j,$$

où  $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j) / \pi_{ij} \pi_i \pi_j$  et  $\pi_{ij}$  est la probabilité d'inclusion de deuxième ordre des unités  $i$  et  $j$ . Notons que  $\pi_{ii} = \pi_i$ . De même, nous avons

$$\tilde{v}_{JS} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \psi_i \psi_j.$$



$$V_{JF} = \sum_{\ell=1}^L \frac{n_{\ell}^g}{n_{\ell}^g - 1} \sum_{j \in s_{\ell}} (Y_{CAL(\ell j)}^* - Y_{CAL}^*)^2. \quad (2.2)$$

L'estimateur de variance  $V_{JF}$  est un estimateur du premier

terme du deuxième membre de (2.1),  $E_p^q V^p(Y_{CAL}^* | \mathbf{r})$ . Ce

terme représente la variance sous le plan de sondage que nous aurions obtenue si les unités répondantes avaient été sélectionnées selon un plan d'échantillonnage aléatoire simple stratifié avec remise, ou, de façon équivalente, si les fractions d'échantillonnage dans les strates,  $(n_h/N_h)$  étaient négligeables. Autrement dit, l'estimateur de variance jackknife complet (2.2) est un estimateur de la variance d'échantillonnage conditionnellement au vecteur des indicateurs de réponse  $\mathbf{r}$ . Par conséquent,  $V_{JF}$  est asymptotiquement sans biais et converge pour  $E_p^q V^p(Y_{CAL}^* | \mathbf{r})$  sous échantillonnage aléatoire simple stratifié avec remise, indépendamment de la validité des hypothèses sous-jacentes. Notons que, puisque  $V_{JF}$  est l'estimateur d'une variance d'échantillonnage, nous pouvons l'obtenir facilement en utilisant un logiciel conçu pour l'estimation jackknife de la variance en présence de données complètes. En d'autres termes, aucun logiciel spécialisé n'est nécessaire. Mentionnons aussi que le deuxième terme du membre de droite de (2.1),  $V^p E_p^q(Y_{CAL}^* | \mathbf{r})$ , n'est pas pris en compte. Donc, l'estimateur de variance jackknife complet ne reflète pas le deuxième terme de (2.1). Toutefois, la contribution de ce terme à la variance totale est négligeable si les fractions d'échantillonnage dans les strates,  $n_h/N_h$ , sont négligeables. Donc,  $V_{JF}$  est asymptotiquement sans biais et converge pour la variance totale,  $V(Y_{CAL}^*)$ . Puisque l'objectif de l'étude est de comparer les estimateurs jackknife complet et simplifié, dans la suite de l'exposé, nous supposons que les fractions d'échantillonnage dans les strates sont négligeables et nous concentrons sur les estimations des totaux, de sorte que nous pouvons omettre l'estimation du deuxième terme de (2.1). Nous constatons que, même si le deuxième terme n'est pas négligeable, nos comparaisons sont valides, car l'estimateur complet et l'estimateur simplifié produisent tous deux une sous-estimation de la variance totale qui correspond au même terme.

Un estimateur de variance jackknife simplifié de  $Y_{CAL}^*$  est donné par

$$V_{JS} = \sum_{\ell=1}^L \frac{n_{\ell}^g}{n_{\ell}^g - 1} \sum_{j \in s_{\ell}} (Y_{CAL(\ell j)}^* - Y_{CAL}^*)^2. \quad (2.3)$$

où  $Y_{CAL(\ell j)}^* = \sum_{h \in s_{\ell}} d_{hi}^{CAL(\ell j)} g_{hi} r_{hi} y_{hi}$ . Notons que les facteurs d'ajustement des poids pour la non-réponse  $g_{hi}$  ne sont pas recalculés dans chaque réplique jackknife. Autrement dit, les facteurs  $g_{hi}$  sont traités comme des constantes, ce qui est inapproprié puisqu'ils dépendent de l'échantillon et de l'ensemble de répondants. Par conséquent, en général,

qui offre une base théorique pour l'étude des propriétés des estimateurs jackknife de variance, peut être décrit comme il suit : premièrement, en appliquant le mécanisme de non-réponse, la population  $U$  est divisée aléatoirement en une population de répondants  $U_r$  et une population de non-répondants  $U_m$ . Puis, sachant  $(U_r, U_m)$ , l'échantillon aléatoire  $s$  est sélectionné conformément au plan d'échantillonnage choisi. La variance totale de  $Y_{CAL}$  peut être exprimée sous la forme

$$V(Y_{CAL}^*) = E_p^q V^p(Y_{CAL}^* | \mathbf{r}) + V^p E_p^q(Y_{CAL}^* | \mathbf{r}), \quad (2.1)$$

où  $E_p^q(\cdot)$  et  $V^p(\cdot)$  désignent l'espérance et la variance par rapport au plan d'échantillonnage et  $E_p^q(\cdot)$  et  $V^p(\cdot)$  désignent l'espérance et la variance par rapport au mécanisme de non-réponse,  $q(\mathbf{r} | \mathbf{I})$ .

À la présente section, nous nous penchons sur l'échantillonnage aléatoire simple stratifié, qui est le plan habituellement utilisé dans les enquêtes-entreprises. Sous ce plan d'échantillonnage, la population  $U$  est partitionnée en  $L$  strates  $U_1, \dots, U_L$  de taille  $N_1, \dots, N_L$ , respectivement. Un échantillon aléatoire simple sans remise  $s_h$  de taille  $n_h$  est tiré de la strate  $h$ ,  $h = 1, \dots, L$ . Chaque échantillon de strate est sélectionné indépendamment et nous supposons que  $n_h \geq 2$  pour tout  $h$ . Dans ce contexte, le poids de sondage de l'unité  $i$  dans la strate  $h$  est  $d_{hi} = N_h/n_h$ . Un estimateur de variance jackknife complet de  $Y_{CAL}$  sous échantillonnage aléatoire simple stratifié s'obtient de la façon suivante :

- i) supprimer l'unité  $(g_j)$  de l'échantillon,  $g = 1, \dots, L$ ,  $j = 1, \dots, n_h$ ;
- ii) ajuster les poids de sondage  $d_{hi}$  pour obtenir les poids jackknife  $d_{hi}^{CAL(\ell j)}$ , où  $d_{hi}^{CAL(\ell j)}$  est donné par
 
$$d_{hi}^{CAL(\ell j)} = \begin{cases} 0 & \text{si } (hi) = (g_j); \\ \frac{n_{\ell}^g}{n_{\ell}^g - 1} d_{hi} & \text{si } h = g, i \neq j; \\ d_{hi} & \text{autrement} \end{cases}$$
- iii) calculer l'estimateur  $Y_{CAL(\ell j)}^*$  de la même façon que  $Y_{CAL}^*$  avec les poids jackknife  $d_{hi}^{CAL(\ell j)}$  au lieu des poids de sondage  $d_{hi}$ ; c'est-à-dire  $Y_{CAL(\ell j)}^* = \sum_{h \in s_{\ell}} w_{hi}^{CAL(\ell j)} r_{hi} y_{hi}$ , où  $w_{hi}^{CAL(\ell j)} = \sum_{i \in s_{\ell}} d_{hi}^{CAL(\ell j)} - \sum_{i \in s_{\ell}} d_{hi}^{CAL(\ell j)} r_{hi} x_{hi}^{CAL(\ell j)}$ ;
- iv) remplacer l'unité supprimée à l'étape (i) dans l'échantillon;
- v) répéter les étapes (i) à (iv) pour toutes les unités  $(g_j)$ ,  $g = 1, \dots, L$ ,  $j = 1, \dots, n_h$ .

Notons que les facteurs d'ajustement pour la non-réponse  $g_{hi}$  sont recalculés dans chaque réplique. Nous obtenons ainsi l'estimateur de variance jackknife complet

d'ajustement de la pondération pour la non-réponse dans une cellule de pondération est calculé en divisant le nombre, pondéré par les poids de sondage, d'unités échantillonnées dans la cellule de pondération par le nombre, pondéré par les poids de sondage, d'unités répondantes dans la cellule de pondération. Nous donnons à cette procédure d'ajustement des poids le nom de procédure d'ajustement par la fréquence. Il s'ensuit que l'estimateur (1.2) se réduit à l'estimateur ajusté par la fréquence.

$$\hat{Y}^{\text{freq.}} = \sum_{c=1}^c \frac{N_c}{N_{rc}} \hat{Y}^{\text{rc.}} \quad (1.4)$$

ou

$$\hat{Y}^{\text{rc.}} = \sum_{i \in s} d_i r_i \delta_{ic} y_i.$$

Le deuxième cas particulier de (1.2) repose sur l'hypothèse qu'une variable continue  $x$  est disponible pour toutes les unités échantillonnées. Soit  $\mathbf{x}_i = (\delta_{i1} x_1, \dots, \delta_{ic} x_i, \dots, \delta_{ic} x_i)'$  et  $\mathbf{z}_i = \delta_i$ . Dans ce cas, le facteur d'ajustement  $g_i$  donné par (1.3) se réduit à  $g_i = X_c / X_{rc} \delta_{ic}$  si l'unité  $i$  appartient à la classe  $c$ , où  $X_c = \sum_{i \in s} d_i \delta_{ic} x_i$  et  $X_{rc} = \sum_{i \in s} d_i r_i \delta_{ic} x_i$ . Ici, le facteur d'ajustement de la pondération pour la non-réponse pour une classe de pondération  $c$  est égal à la somme des données auxiliaires pondérées par les poids de sondage pour les unités dans la cellule de pondération divisée par la somme des données auxiliaires pondérées par les poids de sondage pour toutes les unités répondantes dans la cellule de pondération. Nous donnons à cette procédure d'ajustement des poids le nom de procédure d'ajustement par le ratio. L'estimateur (1.2) se réduit à l'estimateur ajusté par le ratio

$$\hat{Y}^{\text{ratio}} = \sum_{c=1}^c \frac{X_c}{X_{rc}} \hat{Y}^{\text{rc.}} \quad (1.5)$$

Notons que l'estimateur ajusté par la fréquence (1.4) est un cas particulier de l'estimateur ajusté par le ratio quand  $x_i = 1$  pour toutes les unités de population échantillonnées. Enfin, si  $\mathbf{x}_i = \mathbf{z}_i = (\delta_{i1}, \dots, \delta_{ic}, \dots, \delta_{ic} x_i, \dots, \delta_{ic} x_i)'$ , nous obtenons un autre cas particulier de (1.2). Dans ce cas, le facteur d'ajustement  $g_i$  donné par (1.3) se réduit à

$$g_i = N_c \left[ 1 + (\bar{x}_c - \bar{x}_{rc}) \frac{\sum_{i \in s} d_i r_i \delta_{ic} (x_i - \bar{x}_{rc})^2}{(x_i - \bar{x}_{rc})^2} \right],$$

si l'unité  $i$  appartient à la classe  $c$ , où  $\bar{x}_c = X_c / N_c$  et  $\bar{x}_{rc} = X_{rc} / N_{rc}$ . Nous donnons à cette procédure d'ajustement des poids le nom de procédure d'ajustement par *régression linéaire simple*. L'estimateur (1.2) se réduit à l'estimateur ajusté par régression linéaire simple

## 2. Estimation jackknife de la variance

À la section 2, nous discutons des estimateurs de variance jackknife complet et simplifié, et montrons que l'estimateur simplifié est asymptotiquement biaisé. À la section 3, nous évaluons la gravité de ce biais pour deux plans d'échantillonnage utilisés fréquemment. À la section 4, nous présentons les résultats d'une étude par simulation comparant les estimateurs de variance jackknife complet et simplifié. Enfin, nous concluons par certaines observations générales à la section 5.

un modèle de régression logistique. À la section 2, nous discutons des estimateurs de variance jackknife complet et simplifié, et montrons que l'estimateur simplifié est asymptotiquement biaisé. À la section 3, nous évaluons la gravité de ce biais pour deux plans d'échantillonnage utilisés fréquemment. À la section 4, nous présentons les résultats d'une étude par simulation comparant les estimateurs de variance jackknife complet et simplifié. Enfin, nous concluons par certaines observations générales à la section 5.

Les estimateurs (1.4) à (1.6) reposent sur une certaine forme d'ajustement de la pondération dans les classes. Tous sont asymptotiquement sans biais pour  $Y$  si les probabilités de réponse des unités sont égales à l'intérieur des classes (c'est-à-dire que le mécanisme de non-réponse est uniforme à l'intérieur des classes). Cette condition est un cas particulier de la condition (i) discutée plus haut.

Dans le présent article, nous montrons que l'estimateur de variance jackknife simplifié, dans lequel les facteurs d'ajustement sont traités comme étant fixes, a tendance à surestimer la variance réelle de  $Y^{\text{CAL}}$ , au moins dans certains cas simples. Nous prolongeons les travaux de recherche antérieurs de Thompson et Yung (2006) qui ont établi les expressions de la version par linéarisation des estimateurs de variance jackknife complet ainsi que simplifié, et évalué ces expressions empiriquement en utilisant des données provenant de l'Annual Capital Expenditures Survey (ACES) réalisée par le U.S. Census Bureau. Il est intéressant de noter que, dans le contexte de la pondération par la propension à la non-réponse, Kim et Kim (2007) ont montré que traiter les probabilités estimées de réponse comme si elles étaient fixes produit une surestimation de la variance réelle quand les poids d'échantillonnage ne sont pas utilisés dans l'estimation de ces probabilités. Beaumont (2005) a obtenu des résultats comparables dans le contexte de l'imputation quand les probabilités de réponse sont estimées en utilisant un modèle de régression logistique.

$$Y^{\text{regls}} = \sum_{c=1}^c N_c [Y^{\text{rc.}} + (\bar{x}_c - \bar{x}_{rc}) B_{rc}], \quad (1.6)$$

$$B_{rc} = \frac{\sum_{i \in s} d_i r_i \delta_{ic} (x_i - \bar{x}_{rc})^2}{\sum_{i \in s} d_i r_i \delta_{ic} (x_i - \bar{x}_{rc}) (y_i - \bar{y}_{rc})}.$$

ou



où  $w_i = d_i g_i$  et  $g_i$  est un facteur d'ajustement de la pondération pour la non-réponse relié à l'unité  $i$  et donné par

$$g_i = 1 + (\mathbf{X}_i^* - \mathbf{X}_i^*)' \mathbf{T}_i^{-1} \mathbf{z}_i, \quad (1.3)$$

où  $\mathbf{X}_i^* = \sum_{i \in s} d_i r_i \mathbf{x}_i$  et  $\mathbf{T}_i^* = \sum_{i \in s} d_i r_i \mathbf{z}_i \mathbf{x}_i'$ . Quand  $\mathbf{z}_i = \mathbf{x}_i / v_i$ , où  $v_i$  est une constante connue, l'estimateur (1.3) est identique à l'estimateur *Infos* donné dans Särndal et Lundström (2005, équation 7.15). Les propriétés de l'estimateur (1.2) ont été étudiées, entre autres, par Deville (2002), Sautory (2003), Särndal et Lundström (2005), et Kott (2006).

Ici, nous évaluons les propriétés (par exemple biais et variance) de  $Y_{\text{CAL}}$  en utilisant l'approche du modèle de non-réponse sous lequel l'inférence est faite par rapport à la loi conjointe induite par le plan d'échantillonnage et par le mécanisme de non-réponse,  $q(\mathbf{r} | \mathbf{I})$ , où  $\mathbf{I} = (I_1, \dots, I_N)'$  est le vecteur des indicateurs de sélection dans l'échantillon tels que  $I_i = 1$  si l'unité  $i$  est sélectionnée dans l'échantillon et  $I_i = 0$ , autrement et  $\mathbf{r} = (r_1, \dots, r_N)'$  est le vecteur des indicateurs de réponse. Soit  $p_i = P(r_i = 1 | \mathbf{I}, I_i = 1)$  la probabilité de réponse pour l'unité  $i$ . Nous supposons que  $p_i > 0$  pour tout  $i$  et que les unités répondent indépendamment les unes des autres ; autrement dit,  $p_{ij} = P(r_i = 1, r_j = 1 | \mathbf{I}, I_i = 1, I_j = 1, i \neq j) = p_i p_j$ .

L'estimateur  $Y_{\text{CAL}}$  est asymptotiquement sans biais pour le total réel  $Y$  si  $i) p_i^{-1} = 1 + \lambda' \mathbf{z}_i$  pour tout  $i \in U$ , où  $\lambda$  est un vecteur de constantes inconnues ou  $ii) y_i = \mathbf{x}_i' \boldsymbol{\beta}$  pour tout  $i \in U$ , où  $\boldsymbol{\beta}$  est un vecteur de constantes ; voir Särndal et Lundström (2005, chapitre 9.5). Si la condition (i) est satisfaite, l'estimateur ponctuel  $Y_{\text{CAL}}$  est asymptotiquement sans biais pour  $Y$  quelle que soit la variable d'intérêt  $y$  estimée. En outre, il découle de (ii) que  $Y_{\text{CAL}}$  présente un petit biais si les résidus  $E_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$  sont faibles, où  $\mathbf{B} = (\sum_{i \in U} \mathbf{z}_i \mathbf{x}_i')^{-1} \sum_{i \in U} \mathbf{z}_i y_i$ . Par conséquent, le biais de l'estimateur  $Y_{\text{CAL}}$  est petit si le vecteur  $\mathbf{x}$  explique la variable d'intérêt  $y$ . Dans le cas de plusieurs variables d'intérêt, notons que le vecteur  $\mathbf{x}$  pourrait expliquer convenablement une variable d'intérêt donnée, mais ne pas y être relié du tout, auquel cas certaines estimations pourraient être biaisées. Nous supposons que  $Y_{\text{CAL}}$  est asymptotiquement sans biais pour  $Y$ , de sorte que la question du biais des estimateurs étudiés ne se posera pas dans la suite de l'exposé.

Nous considérons trois cas particuliers de (1.2) qui présentent un intérêt pratique (voir également Kalton et Flores-Cervantes 2003). Premièrement, soit  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_N)'$  un vecteur de dimension  $C$  d'indicateurs de classe de pondération attachés à l'unité  $i$  tels que  $\delta_{ic} = 1$  si l'unité  $i$  appartient à la classe  $c$  et  $\delta_{ic} = 0$ , autrement pour  $c = 1, \dots, C$ . Si  $\mathbf{x}_i = \mathbf{z}_i = \boldsymbol{\delta}_i$ , le facteur d'ajustement  $g_i$  donné par (1.3) se réduit à  $g_i = N^c / N_{ic}^c \delta_{ic}$ , où  $N_{ic}^c = \sum_{i \in s} d_i r_i \delta_{ic}$  et  $N_{ic}^c = \sum_{i \in s} d_i r_i \delta_{ic}$ . Autrement dit, le facteur

logistique) ou un modèle non paramétrique (par exemple Da Silva et Opsomer 2006). Un cas particulier de PPN, qui est très répandu en pratique, consiste à d'abord répartir les répondants et les non-répondants en classes de pondération, puis à ajuster les poids de sondage des répondants par l'inverse des taux de réponse dans chaque classe. Ces classes sont formées en se basant sur l'information auxiliaire recueillie pour toutes les unités comprises dans l'échantillon ; voir, par exemple, Eltinge et Vansaneh (1997), ainsi que Little (1986). La deuxième catégorie de procédures d'ajustement, appelée *pondération par calage pour la non-réponse* (PCN), peut être considérée comme une extension de l'approche du calage (Deville et Särndal 1992) adaptée au contexte de la non-réponse totale. Le lecteur est invité à consulter Särndal et Lundström (2005), Kott (2006), ainsi que Brick et Montaquila (2008) pour un survol complet de la PPN et de la PCN. Dans certaines situations, la PPN et la PCN mènent au même estimateur, par exemple, l'estimateur ajusté par la fréquence dans la cellule présente plus loin (voir l'expression (1.4)). Dans le présent article, nous nous concentrons sur la pondération par calage pour la non-réponse (PCN). L'estimation de la variance dans le contexte de la PPN a été étudiée récemment par Kim et Kim (2007).

Considérons une population finie  $U$  de taille  $N$ . L'objectif est d'estimer le total de population  $Y = \sum_{i \in U} y_i$  d'une variable d'intérêt  $y$ . Supposons qu'un échantillon aléatoire  $s$  de taille  $n$  est sélectionné dans  $U$  conformément à un plan donné  $p(s)$ . Dans le cas de données complètes, un estimateur de base de  $Y$  est l'estimateur à facteur d'extension bien connu donné par

$$\hat{Y}_n = \sum_{i \in s} d_i y_i \quad (1.1)$$

où  $d_i = 1/\pi_i$  désigne le poids de sondage relié à l'unité  $i$  et  $\pi_i = P(i \in s)$  désigne la probabilité d'inclusion de premier ordre dans l'échantillon. En présence de non-réponse totale, un sous-ensemble seulement de  $s$  est observé et le calcul de  $\hat{Y}_n$  selon (1.1) est impossible.

Afin de définir un estimateur de  $Y$  ajusté pour la non-réponse, nous supposons qu'il existe un vecteur de variables auxiliaires  $\mathbf{x}$  pour toutes les unités échantillonnées (répondants et non-répondants), de sorte que le vecteur des totaux estimés,  $\mathbf{X}_n^* = \sum_{i \in s} d_i \mathbf{x}_i$ , est disponible. Nous supposons aussi qu'il existe un vecteur de variables instrumentales  $\mathbf{z}$ , un indicateur de dimension que  $\mathbf{x}$ , pour les répondants. Soit  $r_i$  un indicateur de réponse relié à l'unité  $i$  tel que  $r_i = 1$  si l'unité  $i$  est une unité répondante et  $r_i = 0$ , autrement. Pour estimer  $Y$ , nous considérons les estimateurs par calage de la forme

$$Y_{\text{CAL}} = \sum_{i \in s} w_i r_i y_i, \quad (1.2)$$



# L'effet des ajustements pour la non-réponse sur l'estimation de la variance

David Haziza, Katherine Jenny Thompson et Wesley Yung<sup>1</sup>

## Résumé

Dans le cas de nombreux sondages, des procédures d'ajustement des poids sont utilisées pour réduire le biais de non-réponse. Ces ajustements s'appuient sur les données auxiliaires disponibles. Le présent article traite de l'estimation de la variance par la méthode du jackknife pour les estimateurs qui ont été corrigés de la non-réponse. En suivant l'approche inversée d'estimation de la variance proposée par Fay (1991), ainsi que par Shao et Steel (1999), nous étudions l'effet dû au fait de ne pas recalculer l'ajustement des poids pour la non-réponse dans chaque réplique jackknife. Nous montrons que l'estimateur de variance jackknife « simplifié » résultant a tendance à surestimer la variance réelle des estimateurs ponctuels dans le cas de plusieurs procédures d'ajustement des poids utilisées en pratique. Ces résultats théoriques sont confirmés au moyen d'une étude par simulation dans laquelle nous comparons l'estimateur de variance jackknife simplifié à l'estimateur de variance jackknife complet obtenu en recalculant l'ajustement des poids pour la non-réponse dans chaque réplique jackknife.

Mots clés : Calage ; ajustement pour la non-réponse ; non-réponse totale ; estimateur de la variance par la méthode du jackknife ; estimateur de la variance par linéarisation.

## 1. Introduction

La non-réponse totale, c'est-à-dire la situation où, pour une unité échantillonnée, les données manquent pour toutes les variables étudiées ou l'information utilisable n'est pas suffisante, est inévitable dans les sondages. Pour résoudre ce problème, les non-répondants sont supprimés du fichier de données et les poids de sondage des répondants sont ajustés pour tenir compte des unités supprimées. L'objectif principal d'une méthode d'ajustement des poids consiste à réduire le biais de non-réponse qui survient quand les répondants et les non-répondants diffèrent en ce qui concerne les variables étudiées. Pour bien réduire le biais, il est essentiel d'utiliser de l'information auxiliaire puissante, disponible pour les répondants ainsi que les non-répondants.

Dans le présent article, nous considérons l'estimation de la variance par la méthode du jackknife en présence de non-réponse totale. Cette méthode d'estimation de la variance est très répandue en raison de ses propriétés théoriques et de la simplicité des calculs. Contrairement aux méthodes de linéarisation de Taylor, la méthode du jackknife ne nécessite pas le calcul individuel de chaque paramètre d'intérêt ni des probabilités d'inclusion de deuxième ordre qui sont parfois difficiles à obtenir dans les enquêtes complexes. En cas d'utilisation d'un estimateur jackknife de variance dans le contexte de la non-réponse, certains ont soulevé la question de savoir s'il faut ou non produire des répliques de l'ajustement pour la non-réponse (par exemple, Valliant 2004). Dans le présent article, nous considérons deux estimateurs de variance par le jackknife, à

Deux catégories de procédures d'ajustement sont utilisées couramment en pratique. La première, appelée *pondération par la propension* (PPN), consiste à modéliser d'abord les propensions à répondre, puis à utiliser l'inverse des propensions estimées comme facteurs de correction de la pondération. Les propensions à répondre estimées sont habituellement obtenues en ajustant un modèle paramétrique (par exemple, modèle de régression

savoir i) un estimateur de variance jackknife *complet* qui recalculé le facteur d'ajustement pour la non-réponse dans chaque réplique jackknife et ii) un estimateur de variance jackknife *simplifié*, qui ne le fait pas. Ce deuxième estimateur est commode en pratique, mais, autant que nous sachions, ses propriétés théoriques n'ont pas été complètement étudiées dans la littérature. Des considérations tenant à la production ont tendance à dicter l'usage d'un estimateur de variance jackknife simplifié, car dans le contexte de l'échantillonnage stratifié, l'estimateur de variance jackknife complet peut demander assez bien de temps et de ressources informatiques, surtout quand le programme d'enquête utilise un grand nombre de cellules de pondération. Selon certaines études menées récemment par le U.S. Census Bureau (Thompson 2005, ainsi que Ozcoskun, Thompson et Williams 2005), les différences sont négligeables entre les estimations de variance obtenues en utilisant une méthode d'ajustement des poids avec répliques complètes et celles obtenues en utilisant une méthode « simplifiée » avec un estimateur de variance par le jackknife stratifié, le jackknife avec suppression d'un groupe ou la méthode des demi-échantillons modifiés.

<sup>1</sup> David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, H3C 3J7, Canada. Courriel : David.haziza@umontreal.ca ; Katherine Jenny Thompson, U.S. Census Bureau, Washington, DC 20233, Courriel : Katherine.J.Thompson@census.gov ; Wesley Yung, Statistique Canada, Ottawa (Ontario) K1A 0T6, Courriel : wesley.yung@statcan.gc.ca.



- (12) sont utilisés pour générer de nouvelles données latentes  $y^{(i)}$ , sachant la variable binaire observée  $y$  pour l'échantillon, et pour obtenir les valeurs prédites  $\hat{y}^{(i)}$  pour les unités non échantillonnées. Nous pouvons alors obtenir des tirages à partir de la loi a posteriori de la proportion de population finie à l'itération  $i$  sous la forme
- $$\hat{p}_{(i)}^{\text{pr}} = N^{-1} \left( \sum_{i \in S} y_i + \sum_{j \notin S} \hat{y}_j^{(i)} \right).$$
- Albert, J.H., et Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of American Statistical Association*, 88, 669-679.
- Basu, D. (1971). An essay on the logical foundations of survey sampling. Partie 1, dans *Foundations of Statistical Inference* (Eds., V.P. Godambe et D.A. Sprott), Toronto : Holt, Rinehart and Winston, 203-242.
- Compumine (2007). Re: analysis – Tax audit data mining. Février 2007. <http://www.compumine.com/web/public/newstletter/2007/tax-audit-data-mining>.
- Crimmaceanu, C.M., Ruppert, D. et Wand, M. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, 14, 2005, 14.
- Duchesne, P. (2003). Estimation of a proportion with survey data. *Journal of Statistics Education*, 11, 3.
- Eilers, P.H.C., et Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (avec discussion). *Statistical Science*, 11, 89-121.
- Firth, D., et Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Séries B*, 60, 3-21.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3, 515-533.
- Hartley, H.O., et Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- Horvitz, D.G., et Thompson, M.E. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.
- Zheng, H., et Little, R.J.A. (2005). Inference for the population total from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.
- Zheng, H., et Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.
- Yates, F., et Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Séries B*, 15, 235-261.
- Zheng, H., et Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.
- Smith, T.M.F. (1994). Sample surveys 1975-1990: An age of reconciliation? (avec discussion). *Revue Internationale de Statistique*, 62, 5-34.
- Smith, T.M.F. (1976). The foundations of survey sampling: A review (avec discussion). *Journal of the Royal Statistical Society, Séries A*, 139, 183-204.
- Shao, J., et Wu, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.
- Sämdal, C.-E., Swensson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.
- Ruppert, D., Wand, M.P. et Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge, UK : Cambridge University Press.
- Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. et Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Séries B*, 70, 265-286.
- Montanari, G.E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête*, 24, 71-79.
- Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.
- Lehtonen, R., et Veijanen, A. (1998). Estimation de régression généralisées logistiques. *Techniques d'enquête*, 24, 53-58.
- Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2, 813-830.
- Lehtonen, R., Sämdal, C.-E. et Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-673.
- Isaki, C.T., et Fuller, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.



L'estimateur PBSP que nous proposons ici peut être étendu en vue d'inclure des covariables auxiliaires supplémentaires en ajoutant des termes linéaires pour ces variables. Pour l'estimation par domaine, un terme d'interaction entre la fonction spline des probabilités d'inclusion et l'indicateur de domaine devrait également être modélisé. Tant les effets additifs des variables auxiliaires que l'interaction entre l'indicateur de domaine et les probabilités d'inclusion peuvent être représentés dans un modèle mixte (Ruppert et coll. 2003, page 231) et estimés en se servant de l'échantillonnage de Gibbs ou de WinBUGS (Crainiceanu et coll. 2005). L'estimateur PBSP pour les proportions de population finie peut également être étendu à un cas plus général de réponse polychotomique. L'approche de l'échantillonnage de Gibbs pour le cas binaire peut être généralisée au cas des catégories ordonnées et appliquée aux catégories non ordonnées suivant une loi multinomiale latente (Albert et Chib 1993). L'estimateur PBSP peut aussi être étendu à l'estimation sur petits domaines en combinant des effets aléatoires de petit domaine avec la fonction spline lisse sur les probabilités d'inclusion (Opsomer, Claeskens, Kanali, Kauermann et Breidt 2008). Cette extension sera le sujet de futurs travaux de recherche.

Enfin, un examinateur a demandé si l'approche proposée peut être appliquée à une enquête polyvalente comportant de nombreux résultats, puisque la procédure de modélisation ne fournit pas un ensemble unique de poids et doit être répétée pour toutes les variables d'intérêt. Il est vrai que nos méthodes requièrent plus de calculs que les approches existantes, mais la méthode PBSP peut être mise en œuvre facilement avec un algorithme d'échantillonnage de Gibbs ou en utilisant le logiciel WinBUGS, si bien que les calculs ne constituent pas un obstacle majeur. Nous avons mentionné que les simulations décrites dans le présent article comportaient la répétition de l'analyse itérative de Gibbs 6 000 fois, si bien qu'un niveau équivalent de calculs pour une enquête unique de taille comparable permettrait la mise en œuvre de la méthode BPSP pour 6 000 résultats. Ces calculs ont été exécutés sur un PC portable ordinaire. Bien que nous ne défendions l'utilisation automatique d'aucune méthode analytique, fondée sur le plan ou sur un modèle, le fait est que la complexité des calculs n'est plus un obstacle majeur à l'application de ces méthodes. À notre avis, les propriétés statistiques d'une méthode sont plus importantes que le temps de calcul, étant donné les ressources informatiques contemporaines.

## Remerciements

Les présents travaux ont été financés en partie par la société Dow Chemical par la voie d'une subvention sans restrictions accordée pour l'étude sur l'exposition à la

dioxine réalisée par l'Université du Michigan. Les auteurs remercient les examinateurs et un rédacteur associé de leurs commentaires constructifs concernant la version originale du présent article.

## Annexe

### Algorithme de l'échantillonnage de Gibbs

Le modèle (3) peut s'écrire sous forme matricielle,

$$\Phi^{-1}(E(y_i | \beta, b, X, Z)) = (X\beta + Zb), i = 1, \dots, n$$

$$\beta = (\beta_0, \beta_1, \dots, \beta_p)^T, b = (b_1, \dots, b_m) \sim N(0, \tau^2 I_m)$$

$$X = \begin{pmatrix} 1 & \pi_1 & \dots & \pi_p^1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \pi_n & \dots & \pi_p^n \end{pmatrix}, Z = \begin{pmatrix} (\pi_1 - k_1)_p^+ & \dots & (\pi_1 - k_m)_p^+ \\ \vdots & \ddots & \vdots \\ (\pi_n - k_1)_p^+ & \dots & (\pi_n - k_m)_p^+ \end{pmatrix}.$$

L'algorithme de l'échantillonnage de Gibbs pour l'estimation des paramètres du modèle (3) est le suivant :

- le modèle de régression prohibe le résultat binaire  $y = [y_1, \dots, y_n]^T$  correspond à un modèle de régression normal pour des données continues latentes  $y^* = [y_1^*, \dots, y_n^*]^T$ , qui suit une loi normale multivariée tronquée de moyenne  $(X\beta + Zb)$  et de matrice de covariance identité (Albert et Chib 1993), et  $y_i^*$  est l'indicateur que  $y_i^* > 0$ . Partant de certaines valeurs initiales de  $(\beta, b)$ , les valeurs des données continues latentes  $y_i^*$  peuvent être simulées.
- En spécifiant une loi a priori normale aplatie appropriée  $N(0, 10^6)$  sur  $\beta$  et une loi gamma inverse  $GI(0, 1, 0, 1)$  sur  $\tau^2$ , la loi a posteriori de  $(\beta, b, \tau^2)$  sachant les données continues latentes simulées  $y^*$  est

$$(\beta, b) | \tau^2, y^* \sim MVN^{m+p+1}((C^T C + D/\tau^2)^{-1} C^T y^*, (C^T C + D/\tau^2)^{-1})$$

$$\tau^2 | \beta, b \sim GI(0, 1 + m/2, 0, 1 + \|b\|^2/2), \quad (11)$$

où  $C = [X, Z]$  et  $D$  est une matrice diagonale avec  $p + 1$  valeurs de  $10^6$  suivies par  $m$  de ces valeurs sur la diagonale. Gelman (2006) a recommandé d'utiliser une loi a priori uniforme sur  $\tau$ , qui résulte en la loi a posteriori pour  $\tau^2$  de la forme

$$\tau^2 | \beta, b \sim GI(m - 1/2, \|b\|^2/2). \quad (12)$$

- À l'itération  $i$ , des tirages de  $(\beta^{(i)}, b^{(i)}, \tau^{2(i)})$  à partir de la loi a posteriori donnée par l'équation (11) ou

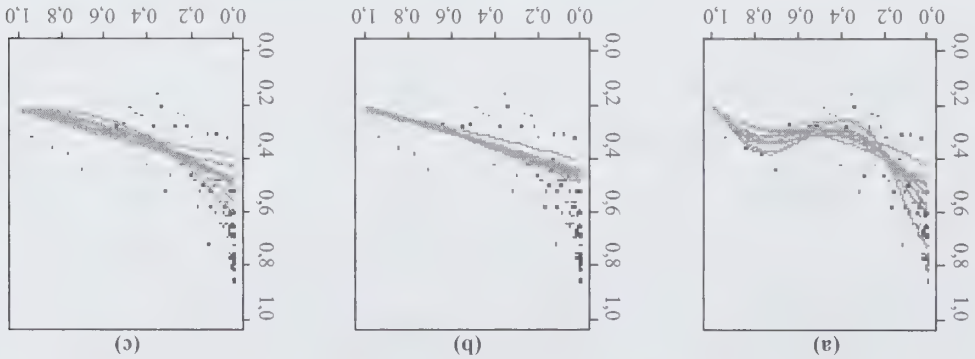


Figure 5 Prédiction fondée sur les données combinées des échantillons pnt et des observations échantillonnées avec certitude dans l'exemple de la vérification fiscale, axe des  $X$  : probabilités réelles  $P(Y=1|\pi)$  dans chaque centile de  $\pi$  ; les points noirs représentent les dix réalisations des moyennes a posteriori de  $P(Y=1|\pi)$ . Les modèles de prédiction sont (a) la régression probit linéaire avec  $p$ -splines, (b) la régression probit linéaire, (c) la régression probit quadratique

Les estimateurs PBSP ne sont pas sensibles aux deux choix de loi a priori pour  $\tau^2$  considérés ici, mais l'exemple de la vérification fiscale semble indiquer que la loi a priori uniforme produit un biais et une REQM un peu plus petits, des intervalles de crédibilité à 95 % plus étroits et une meilleure couverture quant un modèle de prédiction non linéaire est nécessaire. L'exemple de la vérification fiscale montre aussi que, dans l'estimateur RG, une taille de population estimée s'appuyant sur la somme des inverses des probabilités d'inclusion est préférable à la taille réelle de population quand une ou plusieurs observations dont la probabilité d'inclusion est très faible sont incluses dans l'échantillon, puisque l'estimateur RG avec le dénominateur  $N$  possède une variance élevée et une faible efficacité dans ce cas.

Les estimateurs fondés sur le plan et leurs intervalles de confiance à 95 % peuvent fournir des inférences valides pour les proportions de population quand l'échantillon est grand. Cependant, ces propriétés asymptotiques ne semblent pas tenir quand la taille de l'échantillon est moyenne ou faible. L'approche PBSP peut donner des inférences plus valides pour les petits échantillons, surtout quand la proportion de population réelle que l'on veut estimer est proche de 0 ou de 1, quoique la couverture de l'intervalle de confiance semble être inférieure au taux nominal quand la taille d'échantillon diminue et le manque de parcimonie du modèle est un problème. Lors de l'estimation de proportions en dehors des queues de la distribution, l'estimateur PBSP produit une REQM un peu plus faible et une couverture des intervalles de confiance plus proche du taux nominal que les estimateurs HK et RG, mais l'amélioration n'est pas aussi significative que dans les queues. Sous ce scénario, pour

éviter les calculs compliqués de l'estimateur PBSP, l'estimateur PR\_RG basé sur l'équation (7) est une alternative pour les praticiens des sondages.

Le choix de l'estimateur de variance pose un problème pour certains plans de sondage avec probabilités inégales pour les estimateurs fondés sur le plan, mais l'approche de prédiction bayésienne avec  $p$ -splines fournit une approximation par simulation de la loi a posteriori complète de la proportion de population. Un effort supplémentaire en vue d'estimer la variance ou l'intervalle de crédibilité à 95 % de l'estimateur PBSP n'est pas nécessaire, car ceux-ci peuvent être obtenus en même temps que les estimateurs ponctuels. Zheng et Little (2005) comparent trois estimateurs de variance de l'estimateur fondé sur un modèle avec  $p$ -splines pour un total de population finie dans un échantillon pnt, y compris l'estimateur de variance bayésien empirique fondé sur un modèle, l'estimateur de variance jackknife et l'estimateur de variance par la méthode des répliques répétées équilibrées (BRR). Les études en simulation montrent que la méthode du jackknife donne de bons résultats, tandis que la méthode BRR a tendance à produire des erreurs-types prudentes et que l'estimateur bayésien empirique fondé sur un modèle est vulnérable à l'erreur de spécification de la structure de variance. Dans les travaux présentés ici, l'intervalle de crédibilité au niveau  $1 - \alpha$  pour l'estimateur PBSP de la proportion de population est construit en divisant  $\alpha$  également entre les points limites supérieur et inférieur de la loi a posteriori de  $p$ . Cette approche purement bayésienne basée sur des tirages à partir des lois a posteriori semble donner de bons résultats sous les conditions que nous avons établies et évite les calculs lourds associés aux méthodes jackknife et BRR.

Tableau 4  
Comparaison du biais, de la racine carrée de l'erreur quadratique moyenne, ainsi que de la largeur moyenne et du taux de non-couverture des IC à 95 % empiriques de divers estimateurs dans l'exemple des déclarations de revenus

Méthodes	biases*100	REQM*100	largeur moyenne*100	non-couverture*100
HK	-2,4	12,4	10,2	10,2
RL	5,5	11,9	27	45,6
PR	-11,6	12,4	18	83,4
PR_RG1	-1,2	11,5	31	16,8
PR_RG2	-1,2	11,5	33	11,4
PBSP	-6,8	9,3	27	5,0
PBSP_RG1	-3,0	102,6	77	9,2
PBSP_RG2	-0,7	12,0	34	12,8
* RG_1 : estimateurs RG en utilisant l'équation (6); RG_2 : estimateurs RG en utilisant l'équation (7).				

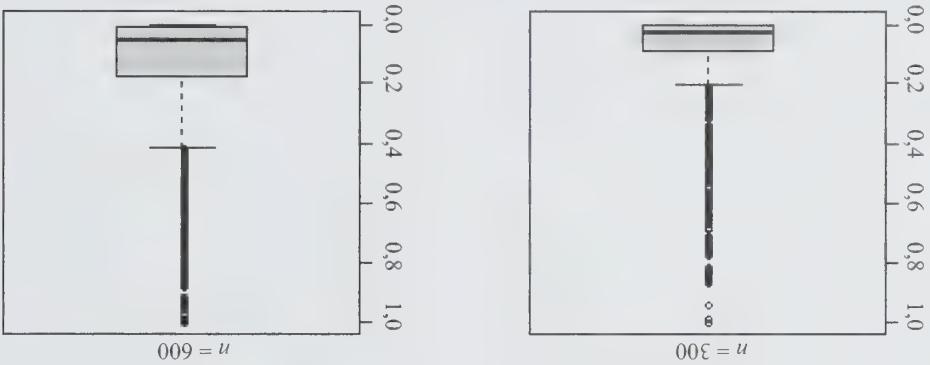


Figure 3 Boîtes à moustache pour les probabilités d'inclusion pour deux tailles d'échantillon dans l'exemple de la vérification fiscale

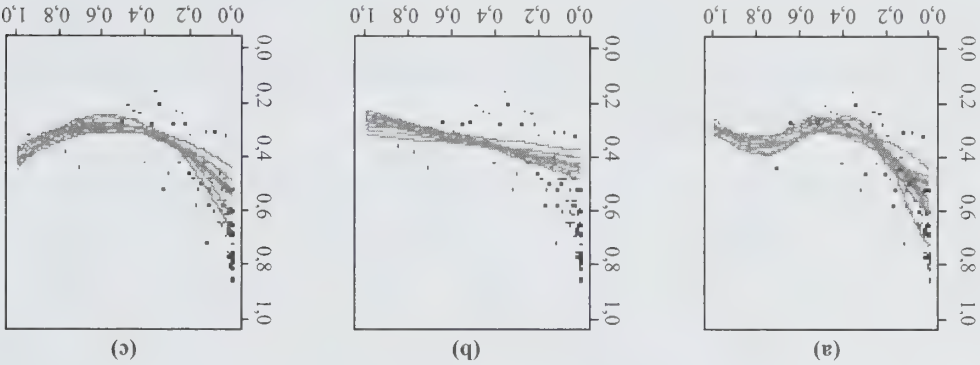


Figure 4 Prédications basées sur les échantillons pnt uniquement dans l'exemple de la vérification fiscale, axe des  $X$  : probabilités d'inclusion  $\pi$ , axe des  $Y$  :  $P(Y=1|\pi)$ ; les points noirs représentent les probabilités réelles  $P(Y=1|\pi)$  dans chaque centile de  $\pi$ ; les courbes grises de prédiction sont (a) la régression probit linéaire avec  $p$ -splines, (b) la régression probit linéaire, (c) la régression probit quadratique



Le tableau 4 montre que l'estimateur PBSP possède un biais légèrement plus grand, mais une REQ<sub>M</sub> plus petite, et un intervalle de crédibilité dont la largeur moyenne est plus étroite et dont la couverture est plus proche du niveau nominal que les estimateurs fondés sur le plan (a), (d) et (f). Des résultats non présentés ici indiquent que l'estimateur PBSP avec une loi a priori uniforme donne d'un peu meilleurs résultats que celui avec une loi a priori gamma inverse en ce qui concerne le biais, la REQ<sub>M</sub> et le taux de couverture empiriques, parce qu'il existe plus de fluctuations dans les données et que la loi a priori uniforme donne plus de flexibilité à la fonction ajustée. L'estimateur PBSP<sub>RG</sub> produit un biais plus faible, mais est moins efficace et a un moins bon taux de couverture que l'estimateur PBSP. L'estimateur de prédiction basé sur le modèle de régression probit linéaire comme modèle de prédiction donne de médiocres résultats ici puisque le modèle est mal spécifié, mais son estimateur RG réduit le biais et la REQ<sub>M</sub>, et améliore le taux de couverture. L'estimateur PBSP<sub>RG</sub> basé sur l'équation (6) donne de très mauvais résultats en ce qui concerne la REQ<sub>M</sub> comparativement à l'estimateur donné par l'équation (7), à cause d'une situation similaire à l'exemple de l'éléphant de cirque de Basu (1971), où une ou plusieurs observations ayant une très faible probabilité d'inclusion sont sélectionnées dans l'échantillon et, donc, reçoivent des poids très grands. Cependant, l'estimateur PR<sub>RG</sub> donne par l'équation (6) a d'aussi bonnes propriétés que celui donné par l'équation (7) avec les prédictions obtenues d'après les estimations pondérées du maximum de vraisemblance, où la probabilité d'inclusion est utilisée comme covariable, ainsi que les poids de sondage. Dans l'ensemble, l'estimateur RG donné par l'équation (7) est préférable à celui donné par l'équation (6). À mesure que la taille d'échantillon augmente, passant de 300 à 600, la probabilité de non-couverture de l'intervalle de crédibilité à 95 % de l'estimateur PBSP s'approche du niveau nominal de 5 % rapidement de 14 % à 5 %, mais les couvertures sont systématiquement inférieures au niveau nominal pour les autres estimateurs.

Comparativement aux estimateurs de prédiction fondés sur un modèle linéaire, l'estimateur PBSP est robuste non seulement à l'erreur de spécification du modèle, mais aussi aux observations influentes présentes dans l'échantillon. Pour démontrer la robustesse aux observations influentes, nous comparons les variations de l'adéquation du modèle en utilisant des modèles probit avec  $p$ -splines, un modèle probit linéaire et un modèle probit quadratique basés sur l'échantillon ppi uniquement à la figure 4, et basés sur l'échantillon ppi ainsi que sur les observations avec probabilités d'inclusion égales à 1 à la figure 5. Dans chaque figure, la population est stratifiée selon les 100 quantiles des probabilités d'inclusion, et les probabilités réelles que  $X = 1$  sont

## 7. Discussion

Les inférences bayésiennes basées sur le modèle avec  $p$ -splines surpassent celles obtenues avec l'estimateur HK, les estimateurs RG et les estimateurs de prédiction fondés sur un modèle linéaire dans nos simulations. Les estimateurs PBSP sont plus efficaces que les estimateurs HK et RG, et, malgré un biais empirique un peu plus grand, la couverture de leurs intervalles de crédibilité à 95 % est meilleure et la largeur moyenne de l'intervalle est plus étroite, surtout quand la proportion de population est proche de 0 ou de 1 et que peu de données provenant des queues de la distribution sont sélectionnées dans l'échantillon. Ces résultats donnent à penser que les travaux de recherche courants sur l'estimation de la prévalence d'événements rares en population finie sont importants. L'estimateur PBSP est une extension naturelle des estimateurs de proportions de population finie fondés sur un modèle de régression linéaire ordinaire. Comparativement aux estimateurs de prédiction fondés sur un modèle linéaire, l'estimateur PBSP est robuste à l'erreur de spécification du modèle et à la présence d'observations influentes dans l'échantillon grâce à l'utilisation d'un modèle avec  $p$ -splines flexible, sans grande perte d'efficacité pour les tailles d'échantillon étudiées. Par conséquent, l'estimateur PBSP est facile à comprendre, mais il requiert des calculs complexes. Toutefois, grâce à WinBUGS, le logiciel statistique bayésien, il peut être implémenté facilement par les praticiens des sondages.

bénéfice positif, les valeurs négatives n'étant pas permises

dans la variable de taille.

Nous avons tiré sans remise à partir de listes de population classées aléatoirement un millier d'échantillons systématiques répétés de taille 300 et 600. Les déclarations contenant les bénéfices les plus importants ont été incluses avec certitude dans les échantillons de taille 300 et 600 : il existait 78 et 241 de ces déclarations, respectivement. La figure 3 montre que la distribution de la probabilité d'inclusion est étalée vers la droite pour la population, même après avoir exclu les observations dont la probabilité d'inclusion est égale à 1.

Nous avons appliqué aux échantillons ppt les 6 mêmes estimateurs que dans l'étude en simulation avec 30 nœuds et comparé leurs propriétés en ce qui concerne le biais, la REQM, ainsi que la largeur moyenne et le taux de non-couverture des intervalles de confiance/crédibilité à 95 % empiriques. Pour l'estimateur PBSP, un nombre fixe de 30 nœuds sont positionnés à des centiles d'échantillon uniformément espacés des probabilités d'inclusion. Pour les estimateurs RG, ni l'estimateur de variance par linéarisation ni celui par le jackknife ne possédant des propriétés essentiellement meilleures que l'autre, nous présentons l'inférence basée sur l'estimateur de variance par linéarisation pour simplifier les calculs. Nous donnons les résultats pour les estimateurs RG basés sur l'équation (6), ainsi que (7) au tableau 4.

REQM empirique  $\times 1\,000$  des six estimateurs (la REQM minimale dans une ligne est en caractères italiques)

Population	<i>n</i>	Prop. réelle	HK	RL	PR	PR_RG	PBSP	PBSP_RG
LINUP	100	0,10	55,1	57,1	46,3	51,3	47,2	51,7
		0,50	65,2	50,8	47,1	49,7	47,7	50,0
		0,90	26,3	22,6	23,3	22,7	23,5	22,9
	200	0,10	39,3	40,9	31,8	36,1	32,0	36,2
		0,50	45,7	35,9	32,8	34,3	32,8	34,6
		0,90	17,8	15,4	15,5	15,4	15,5	15,3
EXP	100	0,10	51,2	60,1	54,4	51,6	51,8	52,4
		0,50	66,1	56,0	43,0	53,2	47,0	51,7
		0,90	24,2	12,4	12,3	12,3	12,3	12,3
	200	0,10	35,9	42,4	39,6	35,6	36,0	36,2
		0,50	45,1	38,9	31,3	36,1	32,1	35,1
		0,90	15,8	8,0	8,1	8,0	8,0	8,0

Tableau 3  
Taux de non-coverage de l'IC à 95 %  $\times$  100 des six estimateurs (le taux de non-coverage le plus proche de 5 dans une ligne est en caractères italique)

Population	<i>n</i>	Prop. réelle	HK	RL	PR	PR_RG	PBSP	PR_RG	VI	V2	PR_RG	PBSP	VI	V2
LINUP	100	0.10	16.2	18.0	8.4	20.9	16.1	9.0	18.4	14.2	7.1	8.4	7.3	14.2
		0.50	7.5	9.4	5.0	7.2	7.6	4.4	7.3	7.1	5.4	8.4	7.1	7.1
		0.90	7.4	11.4	5.7	8.0	9.4	5.4	8.4	7.1	5.4	8.4	7.1	7.1
	200	0.10	10.8	12.6	6.4	13.9	10.9	6.2	12.6	9.4	6.0	9.4	6.0	9.4
		0.50	5.5	8.3	5.5	6.2	5.9	5.1	6.0	5.5	5.5	5.5	5.5	5.5
		0.90	6.0	8.4	4.4	6.1	4.4	4.7	6.3	5.5	5.5	5.5	5.5	5.5
EXP	100	0.10	15.0	18.1	10.5	19.4	14.8	9.2	18.4	14.4	7.2	8.4	7.2	14.4
		0.50	7.4	13.5	12.2	9.0	11.4	8.9	10.2	8.4	7.2	8.4	7.2	8.4
		0.90	6.1	10.5	7.9	9.9	7.6	7.0	9.8	7.2	7.2	7.2	7.2	7.2
	200	0.10	10.8	13.3	9.9	12.5	11.7	7.5	12.4	9.4	7.5	7.5	7.5	9.4
		0.50	6.0	11.5	14.3	7.2	8.5	6.2	7.5	6.9	6.2	7.5	6.2	6.9
		0.90	5.5	8.8	5.5	6.8	4.6	5.5	6.6	3.7	5.5	6.6	5.5	3.7

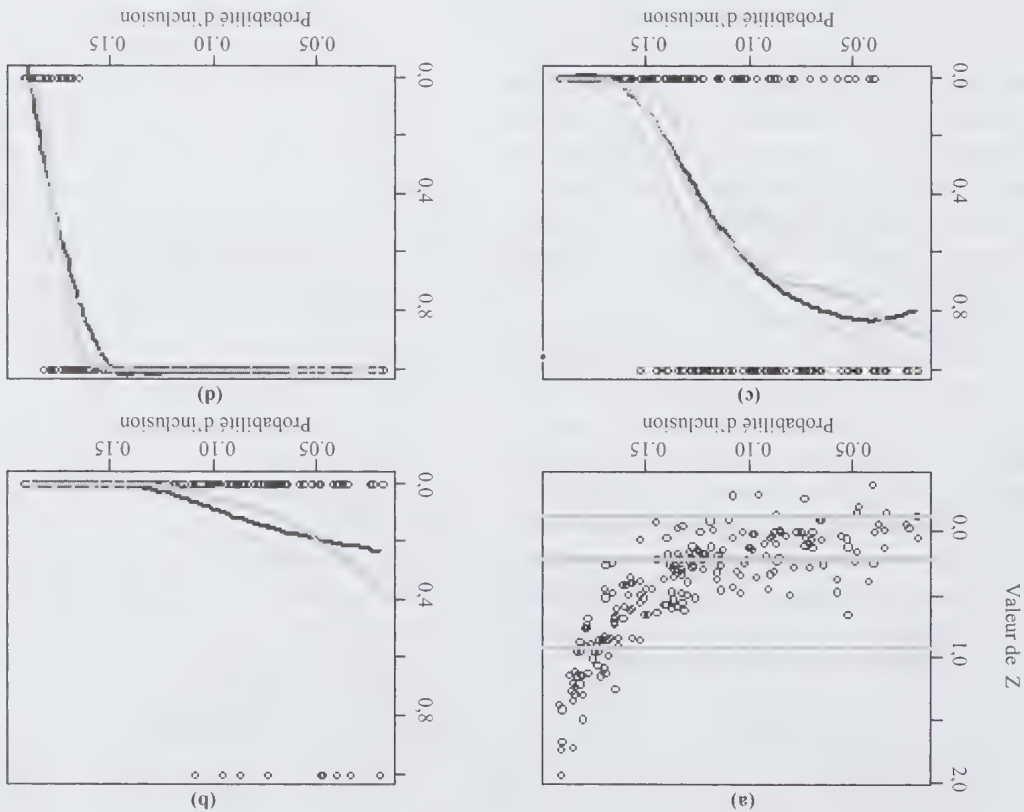


Figure 2 Un échantillon pnt aléatoire tiré du cas EXP ( $n = 200$ ,  $N = 2\,000$ ) : (a) diagramme de dispersion de  $Z$  ; les trois droites grises représentent les 10<sup>e</sup>, 50<sup>e</sup> et 90<sup>e</sup> centiles de la superpopulation, respectivement. (b) Les cercles noirs représentent les unités observées de la variable binaire étudiée  $Y$  dans l'échantillon, définie comme étant  $Y = I(Z \leq 10^{\text{e}} \text{ centile})$  ; les courbes en trait plein et en trait interrompu grises sont les moyennes a posteriori de  $\text{Pr}(Y_i = 1|\pi_i)$  et les intervalles de crédibilité à 95 %, respectivement, simulés en se basant sur un modèle de régression probit avec  $p$ -splines sur  $\pi$  ; la courbe noire est la probabilité  $\text{Pr}(Y_i = 1|\pi_i)$  de superpopulation. (c) Similaire à (b), mais avec  $Y = I(Z \leq 50^{\text{e}} \text{ centile})$ . (d) Similaire à (b), mais avec  $Y = I(Z \leq 90^{\text{e}} \text{ centile})$ .

Tableau 1 Biases empirique  $\times 1\,000$  des six estimateurs (le biais absolu minimum dans une ligne est en caractères italique)

Population	<i>n</i>	Prop. réelle	HK	RL	PR	PR_RG	PBSP	PBSP_RG
EXP	100	0,10	-0,01	13,0	10,3	1,6	8,0	1,2
		0,50	-4,0	-2,9	-4,3	-3,0	-5,2	-3,3
		0,90	-0,4	0,3	-2,5	0,3	-2,9	0,08
	200	0,10	2,5	7,9	5,8	1,5	5,1	1,4
		0,50	3,3	-0,1	-1,3	-0,06	-1,7	-0,2
		0,90	1,6	0,4	-1,0	0,3	-1,2	0,3
EXP	100	0,10	1,2	18,1	25,8	4,7	17,0	3,9
		0,50	-4,0	-3,5	12,5	-1,6	-1,4	-3,4
		0,90	-1,3	-0,2	-1,0	-0,1	-1,0	-0,2
	200	0,10	3,1	11,0	22,1	3,5	13,4	2,7
		0,50	3,8	-0,6	14,0	0,4	0,01	-0,7
		0,90	2,3	0,1	-0,7	0,1	-0,7	0,02

Le choix des priors et hyperpriors dans les modèles mixtes peut avoir une incidence énorme sur les inférences. Nous avons utilisé une loi a priori  $N(0,10^6)$  pour les paramètres à effets fixes,  $\beta_j$ . Dans nos simulations, nous présentons les résultats basés sur une loi a priori gamma inverse appropriée pour  $\tau^2$ , à savoir  $\tau^2 \propto \text{GI}(0,1,0,1)$ . Pour évaluer la sensibilité du choix des lois a priori, nous avons également calculé les résultats en utilisant  $\tau^2 \propto \text{GI}(0,01,0,01)$  et  $\tau^2 \propto \text{GI}(0,001,0,001)$ , ainsi qu'une loi a priori uniforme impropre sur  $\tau$  (Gelman 2006). Ces divers priors ont peu d'effet sur l'inférence a posteriori de la proportion d'intérêt.



courbe en trait plein grise et les deux courbes en trait interrompu grises représentent les moyennes a posteriori de  $\Pr(Y_i = 1 | \pi_j)$  et les intervalles de crédibilité à 95 % basés sur le modèle de régression probit bayésien avec  $p$ -splines linéaires. Les deux autres traces sont similaires au tracé supérieur gauche, mais en prenant les 50<sup>e</sup> et 90<sup>e</sup> centiles de la superpopulation comme valeur seuil pour définir  $Y_i$ . Ces traces montrent que les probabilités réelles que  $Y = 1$  sont comprises dans les intervalles de crédibilité à 95 % et sont proches des moyennes a posteriori de  $\Pr(Y_i = 1 | \pi_j)$ . Nous concluons que le modèle de régression probit bayésien avec  $p$ -splines donne un bon ajustement pour les résultats binaires dans le cas non linéaire.

Le tableau 1 présente le biais empirique ( $\times 10^3$ ) pour les six estimateurs dans les deux populations fondées à partir de LINUP et EXP. Dans l'ensemble, les estimateurs fondés sur le plan de sondage (a, d et f) produisent un biais plus faible que ceux fondés sur le modèle (b, c et e). Dans le cas LINUP, le modèle de régression probit linéaire est spécifié correctement, de sorte que le biais empirique des estimateurs PR est semblable au biais empirique de l'estimateur PBSP ; par contre, dans le cas EXP, les données requièrent l'ajustement d'un modèle de régression probit non linéaire et, donc, le biais de l'estimateur PR est plus grand que celui de l'estimateur PBSP quand les proportions réelles de la population sont 0,1 et 0,5. Cependant, le biais de l'estimateur PBSP à cause de sa propriété de calage interne pour le biais. Comparativement aux estimateurs fondés sur un modèle PR et PBSP, les estimateurs PR\_RG et PBSP\_RG réduisent le biais grâce à l'ajout du terme de calage du biais. En outre, quelque soit le modèle auxiliaire utilisé, les deux estimateurs RG produisent un biais empirique semblable.

Le tableau 2 donne la racine carrée de l'erreur quadratique moyenne empirique ( $\times 10^3$ ) pour les six estimateurs. La racine carrée de l'erreur quadratique moyenne empirique de l'estimateur PBSP est beaucoup plus petite que celle de l'estimateur HK, sauf quand  $p$  est égale à 0,1 dans le cas

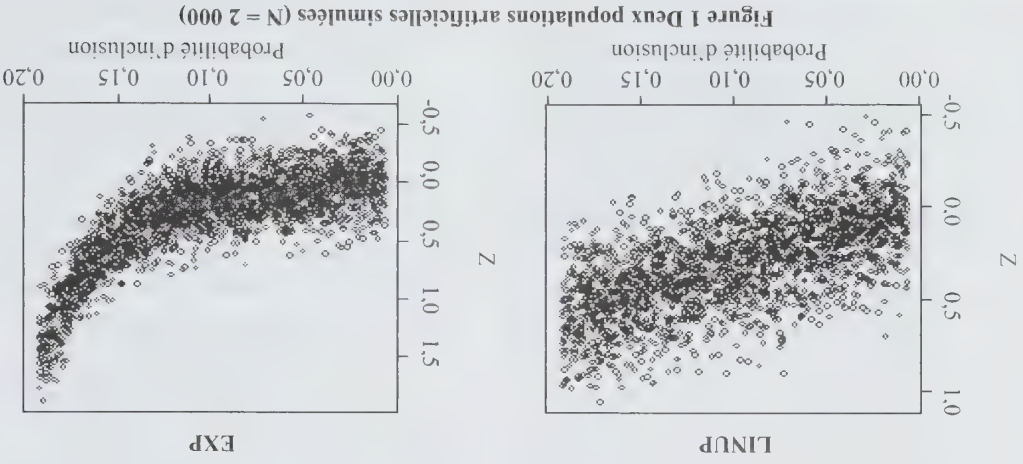


Figure 1 Deux populations artificielles simulées (N = 2 000)

EXP. Dans l'ensemble, l'estimateur PR donne des résultats comparables à l'estimateur PBSP. Afin d'offrir une protection contre l'erreur de spécification du modèle, les estimateurs RG perdent une certaine efficacité comparativement aux estimateurs de prédiction fondés sur un modèle correspondant. L'estimateur PR\_RG a une REQM similaire à l'estimateur PBSP\_RG, mais chacun des deux estimateurs RG a une plus petite REQM que l'estimateur HK grâce à l'utilisation de modèles auxiliaires.

Le tableau 3 donne la probabilité de non-couverture ( $\times 10^3$ ) des intervalles de confiance/crédibilité à 95 %, c'est-à-dire la probabilité que la proportion réelle de population finie se situe à l'extérieur de l'IC à 95 % des estimateurs. Pour calculer les variances des estimateurs, nous utilisons l'estimateur de variance de Yates-Grundy défini par l'équation (2) pour l'estimateur HK, la méthode de rééchantillonnage jackknife définie par l'équation (10) pour l'estimateur RL, ainsi que la méthode de linéarisation (V1) définie par l'équation (9) et la méthode de rééchantillonnage jackknife (V2) pour les estimateurs PR\_RG et PBSP\_RG. Dans l'ensemble, la couverture de l'intervalle de crédibilité est plus proche du taux nominal pour l'estimateur PBSP que pour les cinq autres estimateurs, surtout quand la proportion de population  $p$  est proche de 0 ou de 1, ou que peu d'observations sélectionnées dans l'échantillon proviennent des queues de la distribution. En particulier, l'estimateur PBSP donne lieu à une amélioration importante de la couverture quand  $p$  est proche de 0 tant dans le cas LINUP que dans le cas EXP, puisque peu de données provenant de la queue inférieure des deux populations sont incluses dans l'échantillon. Notons que la couverture améliorée de l'estimateur PBSP est réalisée avec des intervalles qui sont plus étroits en moyenne que ceux des estimateurs HK, LR, PR\_RG et PBSP\_RG. Comme dans le cas du biais et de la REQM empiriques, l'estimateur PBSP\_RG n'améliore pas la couverture comparativement à l'estimateur PR\_RG en utilisant un modèle auxiliaire flexible.

b) RL, estimateur de prédiction de la forme  $\hat{p}_{RL} = N^{-1} (\sum_{j \in S} y_j + \sum_{j \notin S} \hat{y}_{PR}^j)$  avec prédiction  $\hat{y}_{RL}^j$  obtenue en se servant des prédictions du maximum de vraisemblance provenant du modèle de régression logistique linéaire contenant un terme constant et la réciproque de la probabilité d'inclusion comme covariable. RL possède la propriété de « calage interne pour le biais » et est donc convergent sous le plan. RL est exactement le même que son estimateur RG donné par l'équation (6).

c) PR, l'estimateur de prédiction de la forme  $\hat{p}_{PR} = N^{-1} (\sum_{j \in S} y_j + \sum_{j \notin S} \hat{y}_{PR}^j)$  avec la prédiction  $\hat{y}_{PR}^j$  provenant du modèle probit linéaire bayésien contenant un terme d'ordonnée à l'origine et la probabilité d'inclusion comme covariable ;

d) PR<sub>-</sub>RG, l'estimateur RG donné par l'équation (7), où  $\hat{y}_i$  est la prédiction pour l'unité  $i$  quand les paramètres inconnus sont remplacés par les estimations du maximum de vraisemblance pondérées provenant du modèle probit avec un terme constant et la probabilité d'inclusion comme covariable ;

e) PBSP, l'estimateur PBSP défini par l'équation (4), avec  $p = 1$ , une loi a priori gamma inverse pour  $\tau^2$  et l'utilisation de 15 nœuds ;

f) PBSP<sub>-</sub>RG, l'estimateur RG donné par l'équation (7), où  $\hat{y}_i$  est la moyenne a posteriori de  $\Pr(X_i = 1 | \pi_i)$  provenant du modèle PBSP.

Nous donnons uniquement les résultats des simulations basés sur les splines linéaires pour l'estimateur PBSP, puisque les simulations non présentées ici laissent entendre que les splines linéaires donnent d'aussi bons résultats que les splines quadratiques ou les splines cubiques dans tous les scénarios de simulation. Nous avons choisi deux nombres fixes de nœuds (15 ou 30) et avons positionné les nœuds à des centiles d'échantillon uniformément espacés. Les choix de nœuds donnent de bons résultats et un total de 15 nœuds est suffisant pour saisir les courbes dans nos simulations. En outre, les estimateurs RG donnés par (6) donnent des résultats comparables à ceux des estimateurs donnés par (7) ; certaines différences entre ces estimateurs se dégagent dans l'application des données réelles décrites à la section 6, ce qui nous mène à donner la préférence à (7) par rapport à (6).

Nous avons simulé deux populations artificielles de taille 2 000 en utilisant deux lois différentes, avec des taux d'échantillonnage de 5 % et de 10 %, où la variable de taille rend les valeurs entières consécutives 71, 72, ..., 2 070. Nous avons ensuite calculé les probabilités d'inclusion dans la population comme étant proportionnelles à la variable de taille, la valeur maximale correspondant à environ 30 fois les valeurs minimales.

Nous avons d'abord généré des données continues  $Z$  à partir de lois normales ayant une structure de moyenne

## 5.2 Résultats des simulations

$f(\pi)$  et une variance du terme d'erreur constante et égale à 0,04. Nous avons simulé deux structures de moyenne  $f(\pi)$  distinctes, à savoir une fonction linéairement croissante (LINUP)  $f(\pi_i) = k_1 \pi_i$  et une fonction exponentielle (EXP)  $f(\pi_i) = \exp(-4,64 + k_2 \pi_i)$ . Afin que l'étendue de  $Z$  soit la même pour les diverses structures de moyenne,  $k_1$  prend les valeurs de 3 et de 6, et  $k_2$  prend les valeurs de 26 et de 52, quand le taux d'échantillonnage est de 10 % et de 5 %, respectivement. Les deux populations sont représentées graphiquement à la figure 1. Ensuite, nous avons généré la variable de résultat binaire  $X_i$ , qui est égale à 1 si la valeur de  $Z$  est inférieure ou égale à son 10<sup>e</sup> centile de superpopulation, et égale à 0 autrement. De même, nous avons généré les résultats binaires  $Y_2$  et  $Y_3$  en utilisant les 50<sup>e</sup> et 90<sup>e</sup> centiles de  $Z$  en superpopulation comme valeur seuil. Ici, la cible de l'inférence est la proportion de population pour laquelle  $X$  est égale à 1.

Dans chaque réplique simulée, nous avons généré une population finie avant de tirer un échantillon, puis nous avons calculé la proportion de population finie réelle pour laquelle  $X$  est égale à 1 et l'avons désignée  $p$ . Nous avons alors tiré un échantillon ppt systématiquement d'une liste ordonnée aléatoirement de la population et d'échantillon, nous avons obtenu 1 000 répliques et avons comparé les six estimateurs en ce qui concerne le biais, la racine carrée de l'erreur quadratique moyenne (REQM) et le taux de non-couverture de l'intervalle de confiance/crédibilité à 95 % empiriques. Nous présentons les résultats des simulations aux tableaux 1 à 3. Soit  $\hat{p}_i$  une estimation de  $p_i$  basée sur le  $i^{\text{e}}$  échantillon ppt ; le biais et la REQM empiriques sont définis comme il suit :

$$\text{Biais} = \frac{1}{1\,000} \sum_{i=1}^{1\,000} (\hat{p}_i - p_i),$$

$$\text{REQM} = \sqrt{\frac{1}{1\,000} \sum_{i=1}^{1\,000} (\hat{p}_i - p_i)^2}.$$

La figure 2 donne les moyennes a posteriori de  $\Pr(X_i = 1 | \pi_i)$  et les intervalles de crédibilité à 95 % basés sur le modèle probit bayésien avec  $p$ -splines linéaires pour un échantillon ppt aléatoire tiré du cas EXP. Le tracé supérieur gauche est le diagramme de dispersion de la variable continue  $Z$  dans un échantillon ppt, avec trois droites horizontales parallèles superposées représentant les 10<sup>e</sup>, 50<sup>e</sup> et 90<sup>e</sup> centiles de la superpopulation, respectivement. Dans le tracé supérieur droit, la variable binaire  $X_i$  définit le tracé supérieur égal à 1 si la valeur de  $Z$  est inférieure ou égale au 10<sup>e</sup> centile de la superpopulation, est représentée par des cercles noirs, et la  $\Pr(X_i = 1 | \pi_i)$  de superpopulation est représentée par une courbe en trait plein noire. La



propriété est également vérifiée pour les modèles de régression logistique avec  $p$ -splines polynomiales tronquées sur l'inverse des probabilités d'inclusion, ajustés au moyen de la vraisemblance pénalisée. Si nous utilisons la fonction de lien probit au lieu de la fonction de lien logit et que nous effectuons l'ajustement au moyen de l'algorithme de Monte Carlo par chaînes de Markov au lieu du maximum de vraisemblance pénalisée, l'estimateur FBSP pourrait ne plus avoir la propriété de « calage interne pour le biais ». Cependant, la similarité entre le modèle probit et le modèle logistique implique que l'estimateur de prédiction basé sur le modèle de régression probit avec  $p$ -splines est approximativement convergent sous le plan. Nous estimons qu'obtenir des estimations efficaces avec couverture des intervalles de confiance proches du taux nominal dans les échantillons finis est plus important que la convergence exacte sous le plan.

#### 4. Estimateur par la régression généralisée

Pour l'estimation des fréquences de classe d'une variable de réponse discrète, Lehtonen et Veijanen (1998) ont proposé un estimateur par la régression généralisée  $(RG)_{RG}$  du total, qui combine les valeurs prédites  $\hat{y}_i = \Pr(Y_i = 1 | \pi_i)$  en se basant sur un modèle approprié et l'estimateur HT pour les résidus  $r_i' = y_i - \hat{y}_i$  des unités échantillonnées,

$$(5) \quad \sum_{i \in S} \pi_i / \pi_i + \sum_{i \in N} \pi_i = \pi_{RG}$$

L'estimateur RG donné par l'équation (5) est alors utilisé pour construire un estimateur des proportions de population en divisant par la taille connue de population  $N$  (Duchesne 2003).

$$(9) \quad \cdot \left( \sum_{i \in s} \frac{1}{\pi_i} + \sum_{i \in l} \frac{1}{N} \right) \frac{N}{1} = \hat{p}_{RG-1}$$

Nous considérons également ici une autre version de l'estimateur RG pour l'estimation des proportions de population finie, dans lequel le dénominateur du terme de calage du biais pour les résidus  $n_i$  est la taille de population estimée  $\hat{\Sigma}_{n_i} / \pi_i$ .

$$(7) \quad \cdot \left( \sum_{l=1}^{\text{les}} \left( \sum_{l'=1}^{\text{les}} \right) \right) + \sum_{N=1}^{\text{les}} \frac{N}{1} = \hat{p}^{\text{RG}}_{-2} =$$

Pour estimer la variance de (6), nous utilisons l'estimateur de variance du total estimé d'une variable réponse discrète, donné par Lehtonen et Veijanen (1998), divisé par  $N^2$ . Pour estimer la variance de (7), nous appliquons la méthode de linéarisation de Taylor (Särndal, Swensson et Wretman 1992, page 182). Ces deux estimateurs de variance prennent la forme des équations (8) et (9), respectivement,

ment,

Chen, Elliott et Little : inférence basée sur un modèle bayésien avec splines pénalisées

$$(8) \quad \frac{1}{N_2} \sum_{k \in \text{les}} \sum_{l \in \text{res}} \frac{\pi_{kl} \pi_{lk}}{\pi_{ll} \pi_{kk} - \pi_{kl} \pi_{lk}} = \mathcal{V}(\hat{D}_{\text{RG}}^{-1})$$

$$(6) \quad \frac{\pi_l}{e} \frac{\pi_k}{e} \frac{\pi_l}{\pi_l - \pi_k} \sum_{k \in S} \sum_{l \in S} \left( \pi_l / \pi_k \right) = \left( \hat{d}_{\text{RG}}^{-1} \right)$$

où  $e_k = r_k - (\sum_{i \in S} r_i / \pi_i) (\sum_{i \in S} 1 / \pi_i)^{-1}$ . Ces estimateurs de variance requièrent aussi les probabilités d'inclusion par paire, qui peuvent être approchées par la méthode de Hartley et Rao (1962).

Cependant, l'approximation de Hartley et Rao peut donner lieu à un biais dans l'estimateur de variance. Donc, nous considérons également la méthode du jackknife pour l'estimation de la variance (Shao et Wu 1989). Nous stratifions l'échantillon en  $m/G$  strates, chacune de taille  $G$  avec des valeurs similaires des probabilités d'inclusion, puis nous construisons les  $G$  sous-groupes en sélectionnant un élément à la fois dans chaque strate sans remise (Zheng

l'échantillon réduit sans les éléments compris dans le  $\mathcal{G}^o$  sous-groupe, et soit  $\bar{p}$  la moyenne des  $G$  estimateurs basée sur les  $G$  échantillons réduits. L'estimateur de variance jackknife de  $p_{\text{RG}}$  est

$$(10) \quad A^{\text{jackknife}}_{\text{RG}}(\hat{d}) = \hat{d} - \sum_{G=1}^G \frac{G}{G-1} (\hat{d}^{(G)} - \hat{d}).$$

Nous avons utilisé un modèle de régression logistique pondéré par les poids de sondage ajustés sur d'autres covariables comme modèle auxiliaire pour prédire  $\hat{y}_i$  dans les estimateurs RG pour les résultats binaires (Lehtonen et Veijanen 1998 ; Lehtonen et coll. 2005). Comme nous souhitions ici comparer les estimateurs RG avec l'estimateur PBSP, nous appliquons les estimateurs (6) et (7) avec des modèles de régression probit linéaire et des modèles probit avec  $p$ -splines, comme il est décrit en détail à la section 5. Pour l'estimateur RG s'appuyant sur un modèle probit linéaire comme modèle auxiliaire, nous utilisons la probabilité d'inclusion comme covariable, de même qu'un poids dans nos simulations.

## 5. Étude en simulation

## 5.1 Plan de l'étude en simulation

Nous avons réalisé des études en simulation pour étudier la performance de l'estimateur PBSP comparativement à l'estimateur HK, aux estimateurs RG et aux estimateurs de prédiction fondés sur un modèle linéaire pour diverses populations sous échantillonnage ppt. Nous présentons les résultats des simulations pour les six estimateurs suivants :

a) HK, l'estimateur de Hajek défini par l'équation (1) :



égal à  $\{u \times I(n \geq 0)\}^p$  pour tout nombre réel  $u$ . Puisque la fonction de base spline polynomiale tronquée possède  $p - 1$  dérivées continues, des valeurs plus élevées de  $p$  donnent lieu à des splines plus lisses. En spécifiant une loi normale pour  $b$ , l'influence des  $m$  nœuds est contrainte dans le modèle (3), ce qui équivaut à lisser les splines au moyen de la vraisemblance pénalisée.

Les paramètres du modèle (3) peuvent être estimés en utilisant les méthodes s'appliquant au modèle mixte linéaire généralisé. Une autre approche bayésienne qui simplifie les calculs consiste à émettre l'hypothèse de priors et d'hyperpriors faibles et à utiliser l'échantillonnage de Gibbs pour obtenir des tirages à partir des lois a posteriori des paramètres, de la façon suivante : le modèle de régression probit pour les réponses binaires possède une structure de régression normale sous-jacente sur des données continues latentes. Si ces données sont connues, les paramètres des modèles de régression binaire avec  $p$ -splines peuvent être estimés en utilisant les approches classiques pour les modèles de régression normale avec  $p$ -splines. Dans un contexte bayésien, la loi a posteriori des paramètres dans le modèle probit avec  $p$ -splines peut être calculée en utilisant l'échantillonnage de Gibbs (Albert et Chib 1993 ; Ruppert, Wand et Carroll 2003, chapitre 16). En revanche, le modèle de régression logistique avec  $p$ -splines requiert une méthode de calcul plus compliquée, telle que l'algorithme de Metropolis-Hastings. L'avantage en ce qui concerne les calculs rend la fonction de lien probit plus désirable que la fonction de lien logit dans les modèles bayésiens de

pas fournir un ajustement adéquat aux données. Un moyen de résoudre ce problème d'inflexibilité consiste à ajuster une régression binaire sur une fonction spline de  $\pi$  en ajoutant certains nœuds. Toutefois, un trop grand nombre de nœuds peut donner de la « rugosité » à l'ajustement du modèle. Un moyen de surmonter ce problème consiste à garder tous les nœuds, mais à restreindre leur influence en ajustant un modèle de régression binaire avec  $p$ -splines.

Les méthodes courantes de modélisation d'un résultat binaire sont les régressions logistiques et probit, qui donnent généralement des résultats semblables. Nous avons opté pour les modèles probit dans notre étude pour des raisons de commodité des calculs. Le modèle de régression probit pour les résultats binaires possède une structure de régression normale tronquée sous-jacente sur des données continues latentes. Si ces données sont connues, les paramètres des modèles de régression binaire avec  $p$ -splines peuvent être estimés en utilisant les approches classiques pour les modèles de régression normale avec  $p$ -splines. Dans un contexte bayésien, la loi a posteriori des paramètres dans le modèle probit avec  $p$ -splines peut être calculée en utilisant l'échantillonnage de Gibbs (Albert et Chib 1993 ; Ruppert, Wand et Carroll 2003, chapitre 16). En revanche, le modèle de régression logistique avec  $p$ -splines requiert une méthode de calcul plus compliquée, telle que l'algorithme de Metropolis-Hastings. L'avantage en ce qui concerne les calculs rend la fonction de lien probit plus désirable que la

fonction de lien logit dans les modèles bayésiens de régression binaire avec  $p$ -splines.

Les  $p$ -splines peuvent être de divers types. Quand nous appliquons des  $p$ -splines, nous devons choisir leur degré et le positionnement des nœuds, ainsi que les fonctions de base utilisées pour présenter le modèle. Nous choisissons d'utiliser des  $p$ -splines polynomiales tronquées, parce qu'elles sont simples et intuitives. Des estimateurs numériquement plus stables peuvent être obtenus en utilisant des  $B$ -splines par orthogonalisation des bases de fonctions de puissances tronquées (Eilers et Marx 1996). Le modèle de régression probit avec  $p$ -splines polynomiales tronquées se représente comme un modèle mixte linéaire généralisé,

$$\Phi^{-1}(E(y_i | \beta, b, \pi_i)) = \beta_0 + \sum_{p=1}^k \beta_k \pi_i^k + \sum_{l=1}^L b_l (\pi_i - k_l)_+^p \quad (3)$$

$$b_l \sim N(0, \tau^2)$$

$$l = 1, \dots, m; i = 1, \dots, n,$$

où  $\Phi^{-1}(\cdot)$  désigne l'inverse de la fonction de répartition cumulative d'une loi normale centrée réduite, et les constantes  $k_1 < \dots < k_m$  représentent  $m$  nœuds fixes choisis. Une fonction telle que  $(\pi_i - k)_+^p$  est appelée une fonction de base spline polynomiale tronquée de puissance  $p$ , où  $(n)_+^p$  est

L'analogue bayésien d'un intervalle de confiance à  $100 \times (1 - \alpha) \%$  pour la proportion de population est un intervalle de crédibilité à  $100 \times (1 - \alpha) \%$ , qui peut être formé de plusieurs façons. Dans les simulations, nous divisons également l'aire  $\alpha$  située dans la queue entre les points limites supérieur et inférieur.

Firth et Bennett (1998) ont montré que tout modèle de régression logistique paramétrique contenant un terme d'ordonnée à l'origine et l'inverse des probabilités d'inclusion comme covariable, ajusté par le maximum de vraisemblance ordinaire, non pondéré, présente un « calage interne pour le biais » pour les proportions de population et donne donc lieu à la convergence sous le plan. Cette

$$p_{\text{PBSP}} = D^{-1} \sum_{d=1}^J \hat{p}_{\text{PR}}^{(d)} \quad (4)$$

désignée par  $p_{\text{PBSP}}$ , où

simule l'estimateur de prédiction bayésien avec splines pénalisées (PBSP) de la proportion de population finie et est

La loi a posteriori de la proportion de population est bayésienne classique (Crainiceanu, Ruppert et Wand 2005).

binaires d'intérêt et  $p = N^{-1} \sum_{i=1}^N Y_i$ , la proportion de la population pour laquelle  $Y = 1$ . Soit  $\pi_i$  la probabilité d'inclusion de l'unité  $i$ , que l'on suppose connue pour toutes les unités faisant partie de la population finie avant qu'un échantillon soit tiré. Nous tirons alors de la population finie un échantillon aléatoire avec probabilités inégales  $s$  dont les éléments sont  $y_1, \dots, y_n$ , conformément aux probabilités d'inclusion  $\pi_1, \dots, \pi_N$ . L'estimateur HK fondé sur le plan dont il est discuté dans Basu (1971) est défini comme

$$\hat{p}_{HK} = \frac{\sum_{i \in s} y_i / \pi_i}{\sum_{i \in s} 1 / \pi_i}. \quad (1)$$

La variance de  $\hat{p}_{HK}$  peut être estimée par linéarisation de l'estimateur de Yates-Grundy (1953) des totaux,

$$V_{YG}(\hat{p}_{HK}) = \left( \sum_{k \in s} 1 / \pi_k \right)^2.$$

$$V_{YG}(\hat{p}_{HK}) = \left( \sum_{k \in s} 1 / \pi_k \right)^2 \left( \frac{\pi_{i_j}}{\pi_i \pi_j - \pi_{ij}} - \frac{\pi_i}{\pi_j - \hat{p}_{HK}} - \frac{\pi_j}{\pi_i - \hat{p}_{HK}} \right). \quad (2)$$

L'estimateur de variance de Yates-Grundy nécessite les probabilités d'inclusion par paire. Si ces probabilités ne sont pas disponibles, comme cela est le cas dans nos simulations, la formule approximative proposée par Hartley et Rao (1962),

$$\pi_{ij} \approx \frac{n}{n-1} \pi_i \pi_j$$

$$+ \frac{n^2}{n-1} (\pi_i^2 \pi_j + \pi_i \pi_j^2) - \frac{n^3}{n-1} \sum_{k=1}^N \pi_i \pi_j \pi_k^2$$

est souvent utilisée. Un intervalle de confiance de niveau  $1 - \alpha$  approximatif pour la proportion de population  $\hat{p}_{HK}$  est alors obtenu en se basant sur l'approximation normale.

### 3. Estimateur de prédiction bayésien avec splines pénalisées (PBSP)

Royall (1970) a plaidé en faveur de l'utilisation de modèles pour faire des inférences descriptives en population finie en prédisant les valeurs inobservées au moyen de modèles, puisque les inférences fondées sur un modèle devraient être plus efficaces que celles fondées sur le plan de sondage. Afin de modéliser la relation entre le résultat binaire  $Y$  et la probabilité d'inclusion continue  $\pi$ , nous devons ajuster une régression binaire de  $Y$  en  $\pi$ . Les régressions paramétriques binaires, telles que le modèle logistique ou probit, linéaire ou quadratique, pourraient ne

covariable dans le modèle de prédiction (Little 2004). Les estimateurs de prédiction fondés sur un modèle sont convergents et efficaces sous le modèle supposé, mais sont sujets à un biais quand le modèle sous-jacent est mal spécifié. Cette limite motive l'élaboration de modèles statistiques flexibles qui sont plus robustes aux erreurs de spécification du modèle. Pour des données d'enquête continues, Zheng et Little (2003) ont estimé le total de population finie en utilisant une régression non paramétrique sur une fonction spline pénalisée ( $p$ -spline) des probabilités d'inclusion. Nous proposons ici des estimateurs de prédiction bayésiens avec splines pénalisées (PBSP) qui conviennent pour un résultat binaire, par opposition à un résultat continu. Nous adoptons une approche d'inférence bayésienne pour ce modèle, parce que les méthodes bayésiennes produisent souvent une meilleure inférence pour les problèmes avec des petits échantillons et qu'elles peuvent être mises en œuvre de manière commandée pour le modèle que nous proposons au moyen de l'échantillonneur de Gibbs. Dans cette approche, d'autres variables auxiliaires que la probabilité d'inclusion peuvent être incluses dans le modèle, mais nous choisissons la probabilité d'inclusion puisque la modélisation de cette variable est sujette à des erreurs de spécification du modèle.

Nous comparons la performance des estimateurs PBSP à celle des estimateurs de Håjek (HK, type Horvitz-Thompson) et à celle des estimateurs par la régression généralisée (RG) (1998). L'approche par la régression généralisée est une modification populaire de celle assistée par modèle des estimateurs fondés sur le plan de sondage qui consiste à combiner les prédictions issues d'un modèle avec les résidus du modèle pondéré par les poids de sondage (Montanari 1998) pour produire des estimations qui sont approximativement sans biais sous le plan. Zheng et Little (2003 ; 2005) ont comparé par simulation les estimations HT, par prédiction avec  $p$ -splines et RG du total d'une variable observée continue. Ils ont constaté que les estimateurs fondés sur un modèle avec  $p$ -splines donnaient une meilleure erreur quadratique moyenne que les autres méthodes et que les erreurs-types jackknife fournissaient une meilleure couverture des intervalles de confiance que les inférences HT ou RG. Nous procédons à des comparaisons similaires pour l'inférence au sujet d'une proportion de population dans le cas d'un résultat binaire et montrons que notre estimateur PBSP offre les mêmes avantages par rapport aux estimateurs HK et RG.

## 2. Estimateur fondé sur le plan de sondage

Supposons que nous avons une population finie consistée de  $N$  unités identifiables. Soit  $Y$  la variable observée



# Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de population finie dans l'échantillonnage avec probabilités inégales

Qixuan Chen, Michael R. Elliott et Roderick J.A. Little

## Résumé

Nous proposons un estimateur de prédiction bayésien avec splines pénalisées (PBP pour *Bayesian Penalized Spline Predictive*) pour une proportion de population finie sous échantillonnage avec probabilités inégales. Cette nouvelle méthode permet d'intégrer directement les probabilités d'inclusion dans l'estimation d'une proportion de population, en effectuant une régression prohibée du résultat binaire sur la fonction spline pénalisée des probabilités d'inclusion. La loi prédictive a posteriori de la proportion de population est obtenue en utilisant l'échantillonnage de Gibbs. Nous démontrons les avantages de l'estimateur PBP comparativement à l'estimateur de Häjek (HK), à l'estimateur par la régression généralisée (RG) et aux estimateurs de prédiction fondés sur un modèle paramétrique au moyen d'études en simulation et d'un exemple réel de vérification fiscale. Les études en simulation montrent que l'estimateur PBP est plus efficace et donne un intervalle de crédibilité à 95 % dont la probabilité de couverture est meilleure et dont la largeur moyenne est plus étroite que les estimateurs HK et RG, surtout quand la proportion de population est proche de zéro ou de un, ou que l'échantillon est petit. Comparativement aux estimateurs de prédiction fondés sur un modèle linéaire, les estimateurs PBP sont robustes à l'erreur de spécification du modèle et à la présence d'observations influentes dans l'échantillon.

Mots clés : Analyse bayésienne ; données binaires ; régression par splines pénalisées ; probabilité proportionnelle à la taille ; échantillons d'enquête.

## 1. Introduction

Les organismes scientifiques et les administrations publiques utilisent souvent des plans de sondage avec probabilités inégales pour recueillir leurs données. Le plan de sondage avec probabilités inégales le plus simple est sans doute l'échantillonnage stratifié, dans lequel des unités sont échantillonnées dans diverses strates avec des probabilités d'inclusion différentes. Une autre forme importante d'échantillonnage avec probabilités inégales est l'échantillonnage avec probabilités proportionnelles à la taille (ppt), dans lequel la probabilité d'inclusion est proportionnelle à la valeur d'une variable de taille mesurée pour toutes les unités de la population.

Un plan d'échantillonnage avec probabilités inégales tel que l'échantillonnage ppt est fréquemment utilisé pour obtenir des estimations efficaces des moyennes de population de variables continues, pour lesquelles la variance augmente avec la taille de l'unité. Cependant, dans une enquête polyvalente, les inférences au sujet de variables discrètes présentent souvent un intérêt également (par exemple, Lehtonen et Veijanen 1998, Lehtonen, Särndal et Veijanen 2005). Dans le présent article, nous nous attachons aux méthodes d'inférence pour des proportions de population finie sous échantillonnage avec probabilités inégales fondées sur une variable auxiliaire mesurée pour toutes les unités de la population. Nous utilisons l'échantillonnage ppt

comme plan particulier pour illustrer et évaluer nos méthodes.

Les probabilités d'inclusion jouent un rôle important et légèrement différent dans l'inférence fondée sur le plan de sondage et celle fondée sur un modèle en s'appuyant sur des échantillons tirés avec probabilités inégales (Smith 1976, 1994 ; Kish 1995 ; Little 2004). Dans l'inférence fondée sur le plan de sondage, les variables étudiées sont fixes et l'inférence est basée sur la distribution des indicateurs d'inclusion dans l'échantillon ; dans les approches d'estimation classiques fondées sur le plan, telles que l'estimateur de Horvitz-Thompson (HT) (1952) et ses extensions, les unités échantillonnées sont pondérées par l'inverse de leur probabilité d'inclusion. Ces estimateurs sont convergents par rapport au plan (Isaki et Fuller 1982) et fournissent des inférences fiables pour les grands échantillons sans qu'il soit nécessaire de modéliser les hypothèses. Cependant, ces estimateurs peuvent être très inefficaces, comme l'illustre le célèbre exemple de l'éléphant dans Basu (1971). En outre, l'estimation de la variance est fastidieuse, parce qu'elle nécessite les probabilités d'inclusion de deuxième ordre. Les intervalles de confiance correspondants sont fondés sur la théorie asymptotique et peuvent s'écarter des niveaux nominaux pour des tailles d'échantillon moyennes ou faibles.

L'inférence fondée sur un modèle consiste à prédire les valeurs des variables étudiées dans les unités non échantillonnées en introduisant les probabilités d'inclusion comme



- Isakt, C.T., et Fuller, W.A. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.
- Kalton, G., et Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A*, 149, 65-82.
- Lavallée, P. (2007). *Indirect Sampling*. New York : Springer-Verlag.
- Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York : John Wiley & Sons, Inc.
- Lohr, S.L., et Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.
- Lohr, S.L., et Rao, J.N.K. (2006). Estimation in Multiple-frame Surveys. *Journal of the American Statistical Association*, 101, 1019-1030.
- Lu, Y. (2007). Longitudinal estimation in dual frame surveys. Thèse de doctorat, Arizona State University.
- Pfeffermann, D., Skinner, C. et Humphreys, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society, Series A*, 161, 13-32.
- Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Stasny, E.A. (1984). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 25-40.
- Stasny, E.A. (1987). Some Markov-Chain models for nonresponse in estimating gross labor force flows. *Journal of Official Statistics*, 4, 359-73.
- Stasny, E.A., et Fienberg, S.E. (1986). Estimating gross flows using panel data with nonresponse: An example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.
- United States Census Bureau (2006). Current Population Survey: Design and Methodology. Technical Paper 66, U.S. Census Bureau, Washington, DC.
- Verna, V., Betti, G. et Ghellini, G. (2007). Cross-sectional and longitudinal weighting in a rotational household panel: application to EU-SILC. *Statistics in Transition*, 8, 5-50.
- Westat (2001). Survey of Income and Program Participation Users' Guide (Supplement to the Technical Documentation). Rapport technique, Washington, DC.

En utilisant ce modèle, nous pouvons modéliser la non-réponse aux deux périodes (c'est-à-dire perdre à la fois les classifications de ligne et de colonne).

## 6. Conclusion

Dans le présent article, nous avons élaboré des méthodes statistiques pour estimer les flux bruts en nous appuyant sur des enquêtes à base de sondage double. Ces méthodes sont nécessaires pour estimer les changements de situation de pauvreté ou de situation d'emploi au cours du temps. Nous avons élaboré des estimateurs du pseudo-maximum de vraisemblance s'appuyant sur la structure à base de sondage double et les propriétés des deux plans de sondage. Nos modèles tiennent également compte des effets des données manquantes dues au fait qu'une personne cesse de participer à l'enquête ou qu'un plan à panel rotatif est utilisé, de sorte qu'ils permettent d'utiliser pleinement l'information partielle qui peut être fournie par certains ménages. Nous utilisons une méthode jackknife pour estimer la variance des estimateurs et examinons les propriétés de ces derniers. Nous avons appliqué les résultats à des ensembles de données réels.

Dans le présent article, les catégories des tables de contingence des flux bruts sont définies indépendamment des résultats de l'échantillon. Il est également possible de définir les catégories en se basant sur des valeurs qui dépendent de l'échantillon. Par exemple, dans les enquêtes sociales, le seuil de pauvreté pourrait être défini en utilisant un centile fondé sur l'échantillon et les catégories pourraient être définies comme étant « sous le seuil de pauvreté » et « au-dessus du seuil de pauvreté ». Les méthodes exposées dans le présent article peuvent être utilisées pour estimer les flux bruts si les définitions des catégories dépendent de l'échantillon, mais les estimateurs de variance doivent tenir compte de l'effet de l'estimation des bornes des catégories. Bien que les résultats présentés ici aient trait à des enquêtes à base de sondage double, les méthodes sont générales et pourraient être étendues à plus de deux enquêtes en utilisant les estimateurs du pseudo-maximum de vraisemblance (EPMV) élaborés par Lohr et Rao (2006). Toutefois, la complexité des mécanismes éventuels de génération des données manquantes augmente parallèlement au nombre de bases de sondage. Les erreurs de classification pourraient également être plus fréquentes quand le nombre de bases de sondage est plus élevé. Notre étude est effectuée dans le contexte des sondages, mais elle s'applique aussi à d'autres conditions dans lesquelles des données provenant de deux sources indépendantes pourraient être combinées. Comme il devient de plus en plus difficile de couvrir l'entière d'une population d'intérêt au moyen d'une seule enquête, nous pensons que ces

méthodes d'estimation des flux bruts permettent d'obtenir une meilleure couverture de la population à un coût moindre. Elles permettent aussi de compléter une enquête auprès de la population générale par des enquêtes auprès de sous-populations particulières.

## Remerciements

La présente étude a été financée en partie par la National Science Foundation aux termes des subventions SES-0604373 et DLS-0909630. Les auteurs remercient le rédacteur associé et les examinateurs de leurs commentaires avisés et constructifs.

## Bibliographie

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Blair, E., et Blair, J. (2006). Dual frame web-telephone sampling for rare groups. *Journal of Official Statistics*, 22, 211-220.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association*, 63, 542-551.

Catchpole, E.A., et Morgan, B.J.T. (1997). Detecting parameter redundancy. *Biometrika*, 84, 187-196.

Chambers, R.L., Woyzbun, L. et Pillig, R. (1988). Maximum likelihood estimation of gross flows. *Australian Journal of Statistics*, 30, 149-162.

Chen, T., et Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

Fuller, W.A., et Burneister, L.F. (1972). Estimators for samples selected from two overlapping frames. Dans *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.

Hartley, H.O. (1962). Multiple frame surveys. Dans *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā, Series C*, 36, 99-118.

Heeringa, S.G. (1995). Technical description of the assets and health dynamics (ahead) survey sample design. Technical Paper, Institute for Social Research, University of Michigan. [hsronline.isr.umich.edu/docs/userg\\_AHDSAAMP.pdf](http://hsronline.isr.umich.edu/docs/userg_AHDSAAMP.pdf).

Hocking, R.R., et Oxspring, H.H. (1971). Maximum likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, 66, 65-70.

Tableau 5  
Table des flux bruts pour la CPS, en Arizona

Janvier 2002				Janvier 2001			
Données manquantes	Occupée	En chômage	Données manquantes	Occupée	En chômage	Données manquantes	2 607 937
38 848	1 129 656	41 586	57 549	689 497	36 041	606 549	

Puisque nous considérons que la SIPP est un cas sans données manquantes, nous supposons que  $\phi^{kl} = \psi^{kl} = 0$  et utilisons un modèle de type 1 dans l'analyse des données.

Dans les données de la CPS, nous ajustons chaque poids en appliquant le facteur  $2\,625\,091/2\,607\,937$  pour atteindre un total de population unique pour les deux périodes et un total d'observations dans la SIPP (base de sondage 4) après avoir combiné janvier 2001 et janvier 2002 est de 551, et l'effet de plan pour le chômage est de 1,76 environ, de sorte que  $\bar{n}_4 = 551/1,76 = 313$ . L'effet de plan pour le chômage dans la CPS (base de sondage B) est de 1,229 environ, de sorte que  $\bar{n}_B = 1\,020/1,229 = 830$ . En raison des facteurs de vraisemblance, les paramètres estimés des probabilités produits par les cinq modèles données en (3) sont tous les mêmes. Le tableau 6 donne les probabilités estimées et les erreurs-types pour la SIPP, la CPS et les données résultant de la combinaison de ces deux enquêtes.

Tableau 6  
Probabilités de transition estimées en utilisant la SIPP, la CPS et la méthode à base de sondage double avec la SIPP et la CPS. Les erreurs-types sont entre parenthèses

	$P_{00}$	$P_{01}$	$P_{10}$	$P_{11}$
SIPP	0,9489	0,0279	0,0117	0,0115
CPS	(0,0124) 0,9088	(0,0093) 0,0454	(0,0061) 0,0353	(0,0060) 0,0106
SIPP et CPS	(0,0100) 0,9230	(0,0072) 0,0381	(0,0064) 0,0262	(0,0035) 0,0127
	(0,0080)	(0,0058)	(0,0050)	(0,0030)

Pour des raisons de confidentialité, aucune information sur la mise en grappes n'est disponible dans les ensembles de données à grande diffusion de la CPS. Nous avons utilisé le produit de l'effet de plan publié et de la variance pour échantillonnage multinomial pour estimer les variances pour les données de la SIPP, ainsi que de la CPS. Nous avons appliqué le résultat du théorème 1 pour estimer les variances de  $\hat{p}^{kl}$  pour  $k, l = 0, 1$ . Dans cette situation particulière, l'estimation de la variance résultant de la combinaison des deux ensembles de données se réduit à  $(\bar{n}_A/(\bar{n}_A + \bar{n}_B))^2 V_A + (\bar{n}_B/(\bar{n}_A + \bar{n}_B))^2 V_B$ , où  $V_A$  désigne l'estimation de la variance provenant des données de la SIPP et  $V_B$  l'estimation de la variance provenant des données de la CPS. Le tableau 6 montre que les erreurs-types sont réduites si l'on utilise la méthode à base de sondage double.

Nous avons également effectué sur les cinq modèles donnés en (3) les tests d'adéquation élaborés dans Lu (2007). Les estimations des paramètres produites par les cinq modèles et les résultats des tests d'adéquation des modèles sont présentées au tableau 7. Les cinq modèles étant tous bien ajustés aux données, nous recommandons d'adopter le plus simple, c'est-à-dire le modèle 3, pour les données.

Paramètres estimés et résultats des tests d'adéquation				
Paramètres estimés	ddl	$G^2$ corrigé	valeur p	
Modèle 1 $\lambda_{1000}, \lambda_{1100}, \lambda_{1010}, \lambda_{1110}$	3	3,03	0,39	
Modèle 2 $\lambda_1$	5	8,58	0,12	
Modèle 3 $\lambda_0$	5	6,61	0,25	
Modèle 4 $\lambda_{1000}, \lambda_{1100}, \lambda_1$	4	4,10	0,39	
Modèle 5 $\lambda_{100}, \lambda_{1100}, \lambda_{1010}, \lambda_{1110}$	4	6,74	0,15	

Étant donné l'information limitée disponible dans les ensembles de données à grande diffusion, nous avons utilisé des simples corrections des poids pour faire concorder les chiffres estimés de population avec les totaux connus. Les poids inclus dans les ensembles de données de la SIPP et de la CPS ont déjà été calés et corrigés de la non-réponse, de sorte que les modèles utilisés pour les données manquantes reflètent principalement le plan à panel rotatif plutôt qu'une érosion due au démenagement ou à d'autres activités qui pourraient être reliées à la situation d'emploi.

D'autres travaux de recherche sur ces modèles pourraient inclure l'utilisation de diverses corrections de la pondération pour les enquêtes longitudinales. En outre, des paramètres différents pourraient être utilisés pour faire la distinction entre les observations dont la classification est partielle à cause du plan à panel rotatif et celles dont la classification est partielle à cause de la non-réponse. Pour cela, nous pourrions introduire un modèle de chaîne de Markov semblable à celui proposé par Stasny (1987). Dans le modèle avec données complètes, les individus sont répartis dans la table suivant une loi multinomiale unique. À la deuxième étape du processus, qui est également inobservée, chaque individu peut être choisi pour sortir de l'échantillon après l'interview du mois  $t - 1$ , soit entrer dans l'échantillon avant l'interview du mois  $t$ , conformément au plan d'échantillonnage. Enfin, à la troisième étape du processus, chaque individu restant peut soit perdre sa classification de ligne, soit sa classification de colonne pour d'autres raisons.



**Tableau 3**  
Résultats de l'étude en simulations pour les données manquantes générées sous le modèle (1). Le cas (1) correspond à l'ajustement du modèle correct : modèle (1) ; le cas (2) correspond à l'utilisation des registres complets uniquement. Le biais est égal au biais absolu moyen pour les proportions de flux bruts dans la population  $p_{fi}$  ; l'EOME est égal à l'erreur quadratique moyenne empirique moyenne pour les  $p_{fi}$  ; les proportions utilisées pour générer les données manquantes sont  $\lambda_{(r-1)0} = 0,141$ ,  $\lambda_{(r-1)1} = 0,070$ ,  $\lambda_{(r)0} = 0,137$  et  $\lambda_{(r)1} = 0,068$ . Ici,  $n_A$  est le nombre d'UPE dans l'échantillon  $A$  dont la taille est égale à 5 et  $n_B$  est le nombre d'éléments dans l'échantillon  $B$

$n_A$	$n_B$	$P_{00}$			$P_{01}$			$P_{10}$			$P_{11}$		
		Estimateur	Biais	EOME	Estimateur	Biais	EOME	Estimateur	Biais	EOME	Estimateur	Biais	EOME
Cas 1	10	0,311	0,120	0,420	0,311	0,120	0,420	0,311	0,120	0,420	0,311	0,120	0,420
	100	0,040	0,029	0,040	0,040	0,029	0,040	0,040	0,029	0,040	0,040	0,029	0,040
	1000	0,002	0,001	0,002	0,002	0,001	0,002	0,002	0,001	0,002	0,002	0,001	0,002
Cas 2	10	0,286	0,120	0,448	0,286	0,120	0,448	0,286	0,120	0,448	0,286	0,120	0,448
	100	0,048	0,029	0,041	0,048	0,029	0,041	0,048	0,029	0,041	0,048	0,029	0,041
	1000	0,004	0,001	0,002	0,004	0,001	0,002	0,004	0,001	0,002	0,004	0,001	0,002
Cas 1	1000	0,321	0,092	0,449	0,321	0,092	0,449	0,321	0,092	0,449	0,321	0,092	0,449
	1000	0,015	0,011	0,015	0,015	0,011	0,015	0,015	0,011	0,015	0,015	0,011	0,015
	1000	3,337e-04	1,798e-04	3,256e-04	3,337e-04	1,798e-04	3,256e-04	3,337e-04	1,798e-04	3,256e-04	3,337e-04	1,798e-04	3,256e-04
	Estimateur	$\lambda_{r(i)0}$	$\lambda_{r(i)1}$	$\lambda_{r(i)}$	Estimateur	$\lambda_{r(i)0}$	$\lambda_{r(i)1}$	$\lambda_{r(i)}$	Estimateur	$\lambda_{r(i)0}$	$\lambda_{r(i)1}$	$\lambda_{r(i)}$	Estimateur
	Biais	0,145	0,074	0,068	0,145	0,074	0,068	0,145	0,074	0,068	0,145	0,074	0,068
	EOME	2,642e-04	9,389e-05	8,206e-05	2,642e-04	9,389e-05	8,206e-05	2,642e-04	9,389e-05	8,206e-05	2,642e-04	9,389e-05	8,206e-05
100	1000	0,293	0,092	0,480	0,293	0,092	0,480	0,293	0,092	0,480	0,293	0,092	0,480
	Biais	0,0280	0,011	0,040	0,0280	0,011	0,040	0,0280	0,011	0,040	0,0280	0,011	0,040
	EOME	0,001	1,839e-04	0,002	0,001	1,839e-04	0,002	0,001	1,839e-04	0,002	0,001	1,839e-04	0,002
Cas 1	500	0,321	0,093	0,452	0,321	0,093	0,452	0,321	0,093	0,452	0,321	0,093	0,452
	Biais	0,006	0,008	0,012	0,006	0,008	0,012	0,006	0,008	0,012	0,006	0,008	0,012
	EOME	4,960e-05	7,162e-05	1,857e-04	4,960e-05	7,162e-05	1,857e-04	4,960e-05	7,162e-05	1,857e-04	4,960e-05	7,162e-05	1,857e-04
	Estimateur	$\lambda_{r(i)0}$	$\lambda_{r(i)1}$	$\lambda_{r(i)}$	Estimateur	$\lambda_{r(i)0}$	$\lambda_{r(i)1}$	$\lambda_{r(i)}$	Estimateur	$\lambda_{r(i)0}$	$\lambda_{r(i)1}$	$\lambda_{r(i)}$	Estimateur
	Biais	0,140	0,071	0,123	0,140	0,071	0,123	0,140	0,071	0,123	0,140	0,071	0,123
	EOME	4,466e-05	1,818e-05	3,545e-05	4,466e-05	1,818e-05	3,545e-05	4,466e-05	1,818e-05	3,545e-05	4,466e-05	1,818e-05	3,545e-05
500	5000	0,292	0,092	0,483	0,292	0,092	0,483	0,292	0,092	0,483	0,292	0,092	0,483
	Biais	0,028	0,008	0,043	0,028	0,008	0,043	0,028	0,008	0,043	0,028	0,008	0,043
	EOME	8,265e-04	7,642e-05	1,906e-03	8,265e-04	7,642e-05	1,906e-03	8,265e-04	7,642e-05	1,906e-03	8,265e-04	7,642e-05	1,906e-03
Cas 2	500	0,292	0,092	0,483	0,292	0,092	0,483	0,292	0,092	0,483	0,292	0,092	0,483
	Biais	0,028	0,008	0,043	0,028	0,008	0,043	0,028	0,008	0,043	0,028	0,008	0,043
	EOME	8,265e-04	7,642e-05	1,906e-03	8,265e-04	7,642e-05	1,906e-03	8,265e-04	7,642e-05	1,906e-03	8,265e-04	7,642e-05	1,906e-03

## 5. Application

À la présente section, nous appliquons nos résultats à des données provenant de la Survey of Income and Program Participation (SIPP) et de la Current Population Survey (CPS) pour l'Arizona. La CPS et la SIPP sont des enquêtes par panel longitudinales stratifiées à plusieurs degrés. Nous traitons la SIPP et la CPS comme une enquête à base de sondage double ayant la même population cible, à savoir la population de l'Arizona de 18 à 64 ans. En utilisant l'information provenant des deux enquêtes, nous voulons modéliser la variation des probabilités de transition entre les situations d'emploi de janvier 2001 à janvier 2002 chez les personnes de 18 à 64 ans. Notons que, strictement parlant, ces deux enquêtes ne sont pas conçues comme une enquête à base de sondage double. Les questions relatives aux variables de population active ne sont pas les mêmes. Bien que nous ayons recodé les variables conformément aux définitions de la population active appliquées dans la CPS, il

**Tableau 4**  
Table des flux bruts pour la SIPP, en Arizona

Janvier 2002			
(Occupé)		En chômage	
Janvier 2001	Occupé(e)	En chômage	
	2 491 029	30 698	
	73 204	30 160	
	2 625 091		

Pour la CPS, le plan à panel rotatif introduit des données partiellement classifiées. Les mois de janvier 2001 et janvier 2002 ont en commun 50 % de l'échantillon. Nous utilisons ces 50 % de données, ainsi que les données partiellement classifiées pour exécuter l'analyse. La variable de pondération que nous utilisons est un poids transversal avec correction transversales de la non-réponse et calage transversal (United States Census Bureau 2006). Pour les personnes ayant participé à l'enquête l'une des deux années seulement, nous utilisons le poids calculé pour l'année en question. Pour les personnes ayant participé en janvier 2001 ainsi qu'en janvier 2002, nous utilisons la moyenne des deux poids, afin de minimiser la variance de l'estimateur composé. Le groupe de population étudié est celui des 18 à 64 ans, et nous avons exclu les personnes qui n'appartiennent pas à cette catégorie les deux années. Le tableau 5 donne la table de contingence des flux bruts pondérés selon la CPS.

Les deux enquêtes ont pour population cible la population civile à domicile des États-Unis. Nous considérons un sous-ensemble de la population, à savoir la population sur le marché du travail âgée de 18 à 64 ans. Donc,  $N_A = N_B = N_{ab}$  et le problème d'estimation est un cas particulier de la théorie exposée à la section 3. Le fichier longitudinal pour la SIPP de 2001 et de 2002 (Westat 2001) s'appuie sur un seul panel. Nous avons fusionné la vague 1 (contenant les enregistrements de janvier 2001), la vague 4 (contenant les enregistrements de janvier 2002) et le fichier de poids longitudinaux, dans lequel les poids sont corrigés pour que leur somme concorde avec le chiffre de population. Puisque les poids du panel longitudinal ont été corrigés de la non-réponse, nous considérons qu'il s'agit d'un cas sans données manquantes. Le tableau 4 donne la table de contingence résultante des flux bruts pondérés selon la SIPP.

Peut que les différences d'énoncé et d'ordre des questions produisent un biais lorsque l'information est combinée. Nous utilisons ces données comme exemple, parce que des données longitudinales issues d'une base de sondage double réelles ne sont pas disponibles. Néanmoins, l'exemple montre les gains d'efficacité qui peuvent être réalisés en combinant l'information provenant de deux enquêtes pour estimer les flux bruts.

$B$ . Nous avons généré les réponses binaires en grappes pour l'échantillon provenant de la base de sondage  $A$  en créant des vecteurs aléatoires normaux multivariés corrélés, puis en utilisant la fonction `probit` pour convertir les réponses continues en réponses binaires.

Après avoir créé l'échantillon, nous avons calculé les estimateurs des probabilités de l'union de la base de sondage  $A$  et de la base de sondage  $B$ , ainsi que les moyennes des valeurs absolues du biais et des erreurs quadratiques moyennes empiriques (EQME) sous diverses conditions. Nous calculons l'EQME d'un estimateur donné,  $\hat{Y}$ , en nous servant de la formule :

$$\text{EQME} = \frac{1}{R} \sum_{r=1}^R (\hat{Y}_r - Y)^2, \quad (6)$$

où  $\hat{Y}_r$  est la valeur de  $\hat{Y}$  pour la  $r^{\text{e}}$  exécution de la simulation. Dans notre étude en simulations, nous avons utilisé  $R = 100$ .

Nous avons effectué l'étude en simulations en prenant les facteurs suivants : 1)  $\gamma_a : 0,2$  ou  $0,4$ , 2)  $\gamma_b : 0,2$  ou  $0,4$ , 3) paramètre de groupement  $p : 0,3$ , 4) mécanisme de génération des données manquantes : la probabilité qu'une personne soit un non-répondant durant un mois donné dépend de la période et de la classification de la personne durant la période observée, ou données manquantes : près de 10 % ou près de 20 %, 6) tailles de l'échantillon :  $n_A : 10, 100$  ou  $500$ ;  $m : 5$ ,  $n_B : 100, 1\,000$  ou  $5\,000$ . Dans toutes les exécutions, les paramètres de probabilité étaient  $\mathbf{p}_a : (0,3; 0,1; 0,2; 0,4)$ ,  $\mathbf{p}_{ab} : (0,3; 0,1; 0,1; 0,5)$  et  $\mathbf{p}_b : (0,4; 0,1; 0,1; 0,4)$ . Le tableau 3 donne les résultats de l'étude en simulations pour des données manquantes au moyen du modèle 1, quand elles sont prédites au moyen du modèle 1 et au moyen du modèle utilisant les enregistrissements complets uniquement.

Quand les données manquent au hasard, tous les modèles donnent des estimateurs des proportions de flux bruts  $P_{hi}$  approximativement sans biais, si bien que nous ne pré-sentons pas les résultats ici. L'examen du tableau 3 montre que le modèle correct ainsi que l'analyse des enregistrissements complets produisent des estimateurs biaisés des  $P_{hi}$ . Cependant, quand les tailles d'échantillon sont plus grandes, le biais persiste dans l'analyse portant sur les enregistrissements complets uniquement, tandis qu'il diminue quand le modèle 1 est ajusté. Dans l'exemple précédent ici, les probabilités d'avoir des données manquantes sont relativement faibles. Lorsque la quantité de données manquantes est plus importante, le contraste entre les estimateurs est plus prononcé.

$$n_{1/2}^B(\eta_B - \eta_B) \xrightarrow{p} N[0, (1 - (\gamma/(1 + \gamma))) \Sigma_B].$$

Le même argument s'applique pour prouver la convergence et la normalité asymptotique du vecteur d'estimateurs provenant de la base de sondage  $B$ , avec

En combinant ces deux résultats asymptotiques et en utilisant l'indépendance des plans de sondage en même temps que le théorème de Slutsky, nous obtenons (5). La loi limite de  $n_{1/2}^B(\theta - \theta)$  s'ensuit par la méthode  $\Delta$ , puis que les paramètres dans  $\theta$  sont tous des fonctions de  $\eta$ , continuellement dérivables de ceux compris dans  $\eta$ . Puisque les estimateurs des paramètres ne peuvent pas toujours être définis explicitement sous forme d'une fonction d'autres statistiques provenant d'échantillons, nous pouvons dériver les matrices  $\mathbf{H}_A$  et  $\mathbf{H}_B$  en linéarisant les équations de score (Binder 1983). L'hypothèse que  $N^{ab}/N \rightarrow \kappa \in (0, 1)$  garantit que la linéarisation est bien définie.

Le théorème 1 montre que la linéarisation peut être utilisée pour estimer les variances des paramètres d'intérêt. Cependant, dans de nombreuses situations, les matrices  $\mathbf{H}_A$  et  $\mathbf{H}_B$  ont une haute dimensionnalité et les estimateurs de variance linéarisés ont une forme complexe. Un moyen pratique d'estimer les variances des estimateurs consiste à utiliser l'estimateur jackknife proposé par Lohr et Rao (2000). Sous les conditions de régularité exposées dans leur théorème 4, les estimateurs jackknife et par linéarisation de la variance sont asymptotiquement équivalents. La forme de l'estimateur de variance jackknife est  $v_{JK}^K(\theta) = v_A^K(\theta) + v_B^K(\theta)$ , où  $v_A$  est un estimateur jackknife obtenu en sup-primant une unité primaire d'échantillonnage à la fois de la base de sondage  $A$  tout en utilisant l'ensemble de données complet provenant de la base de sondage  $B$ , et  $v_B$  est un estimateur jackknife obtenu en supprimant une unité primaire d'échantillonnage à la fois de la base de sondage  $B$  tout en utilisant l'ensemble de données complet provenant de la base de sondage  $A$ .

## 4.2 Étude en simulations

Le théorème 1 montre que les estimateurs pour base de sondage double sont convergents pour les quantités de population correspondantes sous le mécanisme de génération de données manquantes modélisé. Nous avons exécuté une petite étude en simulations pour examiner les propriétés pour des tailles d'échantillon modérées avec bases de sondage chevauchantes. Nous avons généré les données comme dans l'étude en simulations de Skinner et Rao (1996), avec  $\gamma_a = N_a/N$  et  $\gamma_b = N_b/N$ . Nous avons créé un échantillon en grappes tiré de la base de sondage  $A$  contenant  $n_A$  UPE et  $m$  observations dans chaque UPE, et un échantillon aléatoire simple de  $n_B$  observations pour la base de sondage



## 4. Propriétés des estimateurs

À la présente section, nous examinons les propriétés des estimateurs. Nous calculons les variances asymptotiques, discussions des estimateurs de variance jackknife et exécutons une petite étude en simulations pour explorer les propriétés.

### 4.1 Propriétés

Nous considérons le cas général dans lequel des échantillons stratifiés à plusieurs degrés sont tirés de chaque base de sondage. Les estimateurs des totaux de population ou de Hájek pour les enquêtes complexes. À partir de la base de sondage  $A$ , nous estimons le vecteur de paramètres  $\eta_A = [(\hat{Q}_A^A)', N_A^{ab}/N_A^A]'$  au moyen de  $\hat{\eta}_A = [(\hat{Q}_A^A)', \hat{N}_A^{ab}/N_A^A]'$ , où  $\hat{Q}_A^{klid} = \hat{Q}_A^{klid}/N_A^A$ ; de même, nous estimons  $\eta_B = [(\hat{Q}_B^B)', N_B^{ab}/N_B^B]'$  au moyen de  $\hat{\eta}_B = [(\hat{Q}_B^B)', \hat{N}_B^{ab}/N_B^B]'$  avec  $\hat{Q}_B^{klid} = \hat{Q}_B^{klid}/N_B^B$ .

**Théorème 1 :** Soit  $\hat{\eta} = (\hat{\eta}_A', \hat{\eta}_B')'$  et  $\eta = (\eta_A', \eta_B')'$ . Supposons que  $\hat{\eta}_A$  et  $\hat{\eta}_B$  augmentent tous deux de telle façon que  $\hat{\eta}_A/\eta_B \rightarrow \gamma$  pour une certaine valeur  $0 < \gamma < 1$ . Alors,  $\hat{\eta}$  converge vers  $\eta$ , et

$$\hat{\eta}^{1/2}(\hat{\eta} - \eta) \xrightarrow{d} N(0, \Sigma), \quad (5)$$

où  $\Sigma$  est une matrice diagonale par blocs dont les blocs sont  $\Sigma_A$  et  $\Sigma_B$ ,  $\Sigma_A$  est la matrice de covariance asymptotique de  $\hat{\eta}_A^{1/2}$  et  $\Sigma_B$  est la matrice de covariance asymptotique de  $\hat{\eta}_B^{1/2}$ . Si, en outre, nous supposons que  $N_A^{ab}/N_A \rightarrow \kappa$  pour une certaine valeur  $0 < \kappa < 1$  et que le modèle est identifiable, alors  $\hat{\theta}$  converge vers  $\theta$ , où  $\theta$ , le paramètre d'intérêt, est constitué des composantes de  $p$ ,  $N_A^{ab}/N_A$ ,  $\phi$  et  $\psi$ , et  $\theta$  est l'estimateur du pseudo-maximum de vraisemblance de  $\theta$ . De surcroît,  $\hat{\eta}^{1/2}(\hat{\theta} - \theta)$  est asymptotiquement normal de moyenne 0 et de variance asymptotique  $H_A \Sigma_A H_A' + H_B \Sigma_B H_B'$ , où  $H_F$  est la matrice des dérivées de la fonction  $\theta$  par rapport aux paramètres  $\eta_F$  pour les bases de sondage  $F \in \{A, B\}$ .

**Démonstration.** Dans le cas des flux bruts, les valeurs observées de toutes les variables sont égales à 0 ou 1. Donc, les conditions de bornage figurant dans les lemmes 1 et 2 de Isaki et Fuller (1982) sont satisfaites, et les estimateurs pour la base de sondage  $A$  sont convergents et asymptotiquement normaux avec

$$\hat{\eta}_A^{1/2}(\hat{\eta}_A - \eta_A) \xrightarrow{d} N(0, (\gamma/(1 + \gamma))\Sigma_A).$$

Dans une enquête complexe, particulièrement en cas de mise en grappes, les tailles d'échantillon réelles  $n_A$  et  $n_B$  ne reflètent pas nécessairement les quantités relatives d'information provenant des échantillons. Nous suggérons donc de donner pour valeur à  $\hat{n}_A$  et  $\hat{n}_B$  la taille d'échantillon effective pour chaque échantillon, avec  $\hat{n}_A = n_A/(\text{effet de plan de } S_A)$  et  $\hat{n}_B = n_B/(\text{effet de plan de } S_B)$ . L'effet de plan d'un estimateur  $\hat{\eta}$  est le ratio

$$\frac{[V(\hat{\eta}) \text{ provenant du plan de sondage complexe}]}{[V(\hat{\eta}) \text{ provenant de l'EAS de même taille}]} \\ = \frac{[\hat{n}_A + \hat{n}_B] N_A^{ab, PMV} - [\hat{n}_A N_B^A + \hat{n}_B N_A^A + \hat{n}_A \hat{N}_A^{ab} + \hat{n}_B \hat{N}_B^{ab}] N_A^{ab, PMV}}{[\hat{n}_A \hat{N}_A^{ab} N_B^A + \hat{n}_B \hat{N}_B^{ab} N_A^A] + [\hat{n}_A \hat{N}_A^{ab} N_B^A + \hat{n}_B \hat{N}_B^{ab} N_A^A] = 0.} \quad (4)$$

La pseudo-vraisemblance à la même forme que (2), avec  $\hat{Q}_A^{klid}$ ,  $\hat{Q}_B^{klid}$ ,  $\hat{x}_B^{klid}$ ,  $\hat{x}_A^{klid}$ ,  $\hat{x}_B^{klid}/N_A^A$  et  $\hat{x}_A^{klid}/N_B^B$  remplacés par  $\hat{x}_A^{klid}$ ,  $\hat{x}_B^{klid}$ ,  $\hat{x}_A^{klid}/N_A^A$  et  $\hat{x}_B^{klid}/N_B^B$ . Par suite des facteurs de pseudo-vraisemblance, nous trouvons que  $\hat{N}_A^{ab}$  est égal à la plus petite racine de

L'effet de plan varie habituellement selon la variable. Toutefois, pour estimer les flux bruts, les seuls estimateurs provenant des enquêtes utilisées sont ceux des dénombrements de cellule, et nous pourrions nous attendre à ce que, pour de nombreuses enquêtes, les effets pour les estimateurs  $\hat{x}_A^{klid}$  soient tous les mêmes et qu'ils soient aussi semblables à l'effet de plan de l'estimateur  $\hat{N}_A^{ab}$ . Donc, à l'instar de Skinner et Rao (1996), nous suggérons d'utiliser l'effet de plan pour l'estimateur  $\hat{N}_A^{ab}$  pour déterminer  $\hat{n}_A$ , et l'effet de plan pour l'estimateur  $\hat{N}_B^{ab}$  pour déterminer  $\hat{n}_B$ . Si les effets de plan des autres variables sont effectivement identiques, alors les EPMV résultants minimiseront les variances des quantités estimées ; s'ils diffèrent, les EPMV ne seront pas optimaux, mais ils seront convergents et, dans la plupart des cas, proches des valeurs optimales (Lohr et Rao 2006). Si l'effet de plan pour  $\hat{N}_A^{ab}$  n'est pas connu, comme cela se produirait, par exemple, si l'enquête était poststratifiée en se basant sur  $\hat{N}_A^{ab}$ , alors nous suggérons d'utiliser un effet de plan généralisé, calculé en prenant une moyenne ou une moyenne pondérée des effets de plan d'autres variables de l'enquête.



de la classification de la personne durant la période observée. Sous le modèle 2, la probabilité qu'une personne ne réponde pas à une période donnée dépend uniquement de la période 1. Sous le modèle 3, la probabilité qu'une personne ne réponde pas à une période donnée dépend uniquement de la classification de la personne durant la période observée. Sous le modèle 4, la probabilité qu'une personne ne réponde pas à la période 1 dépend de cette période et de la classification de la personne durant le mois observé. Pour chaque type, de nombreux autres modèles que les cinq susmentionnés sont possibles. En utilisant les matrices des dérivées, il est facile de montrer que les modèles 1 à 5 sont tous identifiables.

En général, il n'existe pas de solution analytique pour les estimations des paramètres, qui doivent donc être obtenues par une méthode itérative. Nous utilisons la fonction « nlm » de R ([www.r-project.org](http://www.r-project.org)) pour calculer les estimations des paramètres ; le code peut être obtenu auprès des auteurs.

### 3.3 Estimateurs pour les échantillons complexes

Quand les données sur l'un des échantillons ou les deux sont recueillies selon un plan de sondage complexe, l'utilisation des dénombrements de cellule directement dans l'expression de la vraisemblance (2) donne des estimateurs qui ne sont pas convergents sous le plan. Skinner et Rao (1996) ont utilisé une méthode du pseudo-maximum de vraisemblance (PMV) pour obtenir des estimateurs convergents sous le plan dans le cas des enquêtes transversales à base de sondage double. Ils ont montré que, contrairement aux estimateurs de Hartley (1962) et de Fuller et Burneister (1972), le même ensemble de poids modifiés était utilisé dans les estimateurs du PMV (EPMV) de diverses variables de réponse et que ces estimateurs étaient donc intérieure-

ment convergents.

Nous proposons d'étudier des estimateurs inspirés de la méthode PMV pour l'estimation des flux bruts dans les enquêtes longitudinales complexes à base de sondage double qui permettent que des données manquent à l'une ou à l'autre période dans l'un ou l'autre échantillon. L'idée fondamentale consiste à émettre l'hypothèse de travail d'une loi multinomiale issue d'une population finie pour donner la forme des estimateurs et à utiliser un effet de plan pour corriger les dénombrements de cellule afin qu'ils reflètent le plan de sondage complexe.

Dans le cas de l'échantillonnage aléatoire simple,  $x_{klt}^A/n_A$  est un estimateur convergent sous le plan de  $Q_{klt}^A$ .

pour la base de sondage  $B$  ont des probabilités non nulles,  $g = q = 16$ . Soit  $D = (D^A, D^B)$  la matrice des dérivées de la transformation, avec  $D_{A(\alpha\beta)}^A = \partial Q_{A(\alpha\beta)}^A / \partial \theta_\beta^A$  et  $D_{B(\beta\gamma)}^B = \partial Q_{B(\beta\gamma)}^B / \partial \theta_\gamma^B$ , pour  $\alpha = 1, \dots, g - 1$ ,  $\beta = 1, \dots, q - 1$ , et  $\gamma = 1, \dots, s$ . Alors, en utilisant les théorèmes 3, 4 et 5 de Carthpole et Morgan (1997), le modèle est localement identifiable si la matrice  $D$  est de plein rang. La preuve pour le cas d'une base de sondage double est donnée dans Lu (2007).

Dans le cas d'une enquête à base de sondage double, nous considérons deux types de modèles pour les données manquantes. Dans un modèle de type (1), la probabilité que l'information manque à la période 1 ou à la période 2 pour la cellule  $(k, l)$  est la même pour chaque domaine dans une base de sondage, c'est-à-dire  $\phi_{klt}^A = \phi_{klt}^B = \phi_{klt}^A = \phi_{klt}^B$ . Dans ce type de modèle, nous estimons les  $\phi$  et les  $\psi$  séparément pour chaque échantillon. Ce modèle pourrait être pris en considération quand les données sur les échantillons provenant des deux bases de sondage sont recueillies en utilisant des modes de collecte différents. Par exemple, si l'échantillon tiré de la base de sondage  $A$  est celui d'une enquête par la poste et l'échantillon tiré de la base de sondage  $B$ , celui d'une enquête par téléphone mobile, on pourrait s'attendre à des probabilités différentes d'abandon pour les deux échantillons.

Dans un modèle de type (2), les probabilités qu'il existe des données manquantes sont les mêmes dans chaque domaine, c'est-à-dire  $\phi_{klt}^A = \phi_{klt}^B = \phi_{klt}^A = \phi_{klt}^B$ . Ce type de modèle pourrait être envisagé si l'on s'attend à ce que la non-réponse soit reliée à l'appartenance à une cellule et que l'on pense que l'appartenance à une base de sondage a peu d'effet sur la non-réponse. Par exemple, si les deux enquêtes s'appuient sur des plans de sondage et des procédures administratives de même type, le choix d'un modèle de type (2) pourrait être approprié.

Pour chaque type de modèle, nous pourrions devoir imposer des contraintes supplémentaires sur les paramètres afin de résoudre les équations de vraisemblance. En nous inspirant de Stasny et Fienberg (1986), les contraintes qui suivent sont possibles :

$$\text{Modèle 1 : } \phi_{klt} = \lambda_{t-1(l)}, \psi_{klt} = \lambda_{t(k)} \quad (3)$$

$$\text{Modèle 2 : } \phi_{klt} = \lambda_{t-1}, \psi_{klt} = \lambda_t$$

$$\text{Modèle 3 : } \phi_{klt} = \lambda_t, \psi_{klt} = \lambda_k$$

$$\text{Modèle 4 : } \phi_{klt} = \lambda_{t-1(l)}, \psi_{klt} = \lambda_t$$

$$\text{Modèle 5 : } \phi_{klt} = \lambda_t, \psi_{klt} = \lambda_{t(k)}.$$

Sous le modèle 1, la probabilité qu'une personne ne réponde pas à une période donnée dépend de cette période et

$\phi^A_{klid}$  de manquer à la période 1 et la probabilité  $\psi^A_{klid}$  de manquer à la période 2. Nous supposons aussi que l'unité ne peut pas manquer aux deux périodes à la fois.

Cette formulation repose sur l'hypothèse que la probabilité qu'une observation manque dans une cellule, un domaine et une base de sondage particuliers est constante. Si des données pouvaient manquer pour d'autres raisons, des paramètres supplémentaires pourraient être utilisés pour faire la distinction entre les observations dont la classification est partielle, à cause, disons, du plan à panel rotatif et celles dont la classification est partielle à cause de la non-réponse. À la section 5, nous discutons d'une autre approche qui pourrait être utilisée avec des mécanismes multiples de génération des données manquantes.

Pour  $k, l \in \{0, 1\}$ , la probabilité qu'une unité provenant de  $S_A$  soit observée dans la cellule  $(k, l)$  et le domaine  $d$  est

$$\tilde{O}^A_{klid} = P^A_{klid}(1 - \phi^A_{klid} - \psi^A_{klid}).$$

La probabilité qu'une unité provenant de  $S_A$  soit observée dans la cellule  $(k, l)$  et le domaine  $d$  est

$$\tilde{O}^A_{klid} = \sum_{l=0}^1 P^A_{klid} \psi^A_{klid}.$$

De même, la probabilité qu'une unité provenant de  $S_A$  soit observée dans la cellule  $(M, l)$  et le domaine  $d$  est

$$\tilde{O}^A_{Mlid} = \sum_{k=0}^K P^A_{klid} \phi^A_{klid}.$$

Nous définissons de la même manière les probabilités pour la base de sondage  $B$ , soit  $\tilde{O}^B_{klid} = P^B_{klid}(1 - \phi^B_{klid} - \psi^B_{klid})$ ,  $\tilde{O}^B_{klid} = \sum_{l=0}^1 P^B_{klid} \psi^B_{klid}$  et  $\tilde{O}^B_{Mlid} = \sum_{k=0}^K P^B_{klid} \phi^B_{klid}$ .

Sous ce modèle à deux phases et en émettant l'hypothèse d'indépendance des échantillons, la fonction de vraisemblance pour les deux échantillons est donnée par :

$$L(\mathbf{p}, \psi, \phi, N^{ab}) \propto \prod_{k \in \{0, 1\}} \prod_{l \in \{0, 1\}} \prod_{d \in \{a, b\}} \prod_{i \in \{0, 1\}} \prod_{j \in \{0, 1\}} \prod_{d \in \{a, b\}} (\tilde{O}^A_{klid})^{x^A_{klid}} \times \prod_{k \in \{0, 1\}} \prod_{l \in \{0, 1\}} \prod_{d \in \{a, b\}} (\tilde{O}^B_{klid})^{x^B_{klid}} \times \prod_{k \in \{0, 1\}} \prod_{l \in \{0, 1\}} \prod_{d \in \{a, b\}} (\tilde{O}^A_{Mlid})^{x^A_{Mlid}} \times \prod_{k \in \{0, 1\}} \prod_{l \in \{0, 1\}} \prod_{d \in \{a, b\}} (\tilde{O}^B_{Mlid})^{x^B_{Mlid}}, \quad (2)$$

où  $\psi$  est le vecteur des  $\psi^A_{klid}$  et des  $\psi^B_{klid}$ , et  $\phi$  est le vecteur des  $\phi^A_{klid}$  et des  $\phi^B_{klid}$ .

L'expression (2) correspond au modèle le plus général, dans lequel les deux enquêtes sont longitudinales et présentent toutes deux des données manquantes à chaque période. Si la base de sondage  $A$  est utilisée dans une enquête à panel rotatif, par exemple, alors toutes les probabilités  $\tilde{O}^A_{klid}$  sont non nulles : les unités comprises dans les panels mesurés aux deux périodes seront incluses dans les estimateurs  $x^A_{klid}$  pour  $k, l \in \{0, 1\}$ , les unités dans les panels sortant de l'enquête après la période 1 seront incluses dans les estimateurs  $x^A_{klid}$  et les unités figurant dans les panels entrants seront incluses dans les estimateurs  $x^A_{klid}$ . Selon la structure des enquêtes, certains facteurs de l'expression (2) peuvent être omis. Par exemple, si l'enquête s'appuyant sur la base de sondage  $B$  est une enquête transversale répétée dont la traction d'échantillonnage est faible, les probabilités  $\tilde{O}^B_{klid}$  pour  $k, l \in \{0, 1\}$  seront presque nulles et nous omettrons ces facteurs dans l'expression de la vraisemblance.

La vraisemblance donnée par (2) peut s'écrire sous la forme du produit d'un facteur contenant  $N^{ab}$  et d'un facteur contenant les autres paramètres. Par conséquent, l'EMV pour  $N^{ab}$  est de nouveau la plus petite racine de l'équation (1). Nous discutons des estimateurs des paramètres restants à la section suivante.

### 3.2.2 Identifiabilité du modèle et modèles réduits

L'une des difficultés que pose la maximisation de la vraisemblance donnée par (2) est que, sous le modèle général, il existe un total de 42 paramètres, tandis que les deux échantillons ne comprennent que 32 dénombrements de cellule observés. Donc, nous ne pouvons pas estimer tous les paramètres sous le modèle le plus général. Toutefois, nous pouvons envisager des modèles dont le nombre de paramètres est réduit, comme l'ont fait Chen et Fienberg (1974) pour les enquêtes à base de sondage unique. En fait, le cas de la base de sondage double donne nettement plus de souplesse pour la modélisation des données manquantes, grâce à l'information indépendante provenant des deux échantillons au sujet du domaine  $ab$ .

Nous commençons par énoncer les conditions pour qu'un modèle réduit soit localement identifiable. Soit  $\theta$  le vecteur de dimension  $s$  des paramètres d'intérêt ; dans notre cas,  $\theta$  comprend les composantes linéairement indépendantes de  $\mathbf{p}$ ,  $N^{ab}/N$  et les paramètres pour le mécanisme de génération des données manquantes. Dans l'expression de la vraisemblance (2), les probabilités provenant des échantillons multinomiaux indépendants sont  $\tilde{O}^A_{klid}$  et  $\tilde{O}^B_{klid}$ . Ces probabilités pourraient s'écrire sous la forme de fonctions de  $\theta$ , avec  $\tilde{O}^A_{klid}(\theta) = (\tilde{O}^A_{klid}{}^{00ab}, \dots, \tilde{O}^A_{klid}{}^{11ab})$ , un vecteur de dimension  $g$  des  $\tilde{O}^A_{klid}$  non nulles et  $\tilde{O}^B_{klid}(\theta) = (\tilde{O}^B_{klid}{}^{00ab}, \dots, \tilde{O}^B_{klid}{}^{11ab})$ , un vecteur de dimension  $q$  des  $\tilde{O}^B_{klid}$  non nulles. Quand toutes les cellules du tableau 2 et celles du tableau analogues



enquêtes à base de sondage double. Nous suivons une approche de pseudo-vraisemblance à base de sondage double pour tenir compte des plans d'échantillonnage et des mécanismes de génération des données manquantes. L'approche à base de sondage double permet d'améliorer la précision des estimateurs et offre plus de souplesse pour modéliser les mécanismes susmentionnés. Les méthodes utilisées à l'heure actuelle pour traiter les données manquantes sont fondées sur des méthodes statistiques classiques et entrent dans quatre catégories générales (Little et Rubin 2002) : l'analyse des cas complets, les méthodes de pondération, les méthodes d'imputation et les méthodes fondées sur un modèle. Ici, nous adoptons une approche fondée sur un modèle pour traiter les données manquantes. À la présente section, nous considérons des conditions simples, c'est-à-dire des échantillons aléatoires simples tirés d'une population sans données manquantes. Puis, nous ajoutons un modèle pour le mécanisme de génération de données manquantes. Enfin, nous discutons d'estimateurs applicables à des plans de sondage plus complexes.

### 3.1 Échantillons aléatoires simples avec données complètes

Pour justifier l'estimateur utilisé dans le cas général, nous commençons par étudier l'estimation des flux bruts quand il n'existe pas de données manquantes et quand l'échantillon tiré de chaque base de sondage est un échantillon aléatoire simple. Alors,  $x_{kl}^A = n_A X_{kl}^A / N_A$ , pour  $d = a, ab$  est le nombre observé d'unités échantillonnées dans la cellule  $kl$  et le domaine  $d$  provenant de  $S_A$  ;  $x_{kl}^B = n_B X_{kl}^B / N_B$  pour  $d = b, ab$  est le nombre observé correspondant d'unités échantillonnées provenant de  $S_B$ .

Si les fractions d'échantillonnage sont faibles, nous pouvons utiliser une approximation multinomiale de la vraisemblance. Dans le cas de l'échantillon tiré de la base de sondage  $A$ , il existe huit cellules dont les probabilités associées sont  $P_A^{kl} = p_A^{kl} N_A^d / N_A$ , pour  $k, l \in \{0, 1\}$  et  $d \in \{a, ab\}$ . Les probabilités correspondantes pour l'échantillon tiré de la base de sondage  $B$  sont  $P_B^{kl} = p_B^{kl} N_B^d / N_B$  pour  $k, l \in \{0, 1\}$  et  $d \in \{b, ab\}$ . En utilisant la loi multinomiale et en supposant que les échantillons provenant des deux bases de sondage sont tirés indépendamment, la fonction de vraisemblance est donnée par

$$L(\mathbf{p}, N^{ab}) \propto \prod_{k,l,d} (P_A^{kl})^{x_{kl}^A} \times \prod_{k,l,d} (P_B^{kl})^{x_{kl}^B}.$$

Pour simplifier, nous écrivons la vraisemblance en fonction de  $P_A^{kl}$  et  $P_B^{kl}$ , mais les paramètres d'intérêt sous-jacents sont  $\mathbf{p} = (P_{00a}, P_{01a}, \dots, P_{11b})$  et  $N^{ab}$ .

En posant que les dérivées partielles de la log-vraisemblance par rapport aux paramètres sont nulles, les estimateurs du maximum de vraisemblance sont donnés par  $\hat{p}_{kl}^A = x_{kl}^A / n_A^A$ ,  $\hat{p}_{kl}^B = x_{kl}^B / n_B^B$  et  $\hat{p}_{kl}^{ab} = (x_{kl}^A + x_{kl}^B) / (n_A^A + n_B^B)$ .

où  $n_A^{ab} = \sum_{j \in S_B} I(j \in ab)$ ,  $n_B^{ab} = \sum_{j \in S_A} I(j \in ab)$ ,  $n_A^A = n_A^A - n_B^A$  et  $n_B^B = n_B^B - n_A^B$  pour  $N^{ab}$ ,  $\hat{N}^{ab}$ , est la plus petite racine de l'équation quadratique

$$[n_A^A + n_B^B] \hat{N}^2 - [n_A^A n_B^B + n_B^A n_A^A + n_B^B n_A^A + n_B^A n_B^B] \hat{N}^{ab} + [n_A^A n_B^B + n_B^A n_A^A] N^A N^B = 0. \quad (1)$$

Enfin, en utilisant les résultats susmentionnés, nous construisons les EMV pour  $X_{kl}^A$  et  $P_{kl}^A$  :

$$\hat{P}_{kl}^A = \frac{(N_A^A - \hat{N}^{ab}) \hat{P}_{kl}^A + \hat{N}^{ab} \hat{P}_{kl}^{ab} + (N_B^B - \hat{N}^{ab}) \hat{P}_{kl}^B}{(N_A^A - \hat{N}^{ab}) \hat{P}_{kl}^A + \hat{N}^{ab} \hat{P}_{kl}^{ab} + (N_B^B - \hat{N}^{ab}) \hat{P}_{kl}^B}.$$

Ces estimateurs sont les mêmes que ceux obtenus par Skinner (1991). Cependant, ce dernier a utilisé la loi normale approximative de la moyenne des réponses  $\bar{y}$  dans chaque domaine pour obtenir les EMV, tandis que nos estimateurs proviennent d'un modèle multinomial. Ce modèle nous permet d'inclure des données partiellement classifiées provenant des unités observées durant une seule période, comme nous le montrons à la section suivante.

### 3.2 Échantillons aléatoires simples avec données manquantes

En pratique, certaines personnes peuvent n'apparaître dans l'échantillon qu'à l'une des deux périodes seulement. Cette situation peut être due à l'érosion de l'échantillon (quand des membres de l'échantillon cessent de participer à l'étude) ou à d'autres causes. Dans une enquête à panel rotatif, telle que la CPS, les personnes qui sortent de l'échantillon à la période 1 ne seront pas interrogées à la période 2 et, par conséquent, leur situation d'emploi durant cette période sera inconnue. Dans d'autres situations, l'un des échantillons peut être transversal, auquel cas toutes les observations sont mesurées exactement à une seule période.

#### 3.2.1 Modèle pour les données manquantes

Blumenthal (1968), Chen et Fienberg (1974), Stasny (1984, 1987), et Stasny et Fienberg (1986) ont utilisé une procédure à deux phases pour modéliser les données manquantes dans un échantillon unique. Un modèle est d'abord proposé pour les données complètes, puis le mécanisme de génération des données manquantes est modélisé. Nous étendons cette procédure à nos structures à base de sondage double. L'un des avantages d'une enquête à base de sondage double est qu'elle offre plus de souplesse pour la modélisation des données manquantes.

Premièrement, nous supposons que si toutes les unités étaient mesurées aux deux périodes, le modèle de la section 3.1 pourrait être utilisé. Pour le mécanisme de non-réponse, nous supposons que chaque observation dans la cellule  $(k, l)$  et le domaine  $d$  provenant de  $S_A$  a la probabilité



Population Survey et à la Survey of Income and Program Participation. Enfin, à la section 6, nous présentons nos conclusions.

2. Notation et quantités dans les échantillons

Supposons qu'il existe deux bases de sondage,  $A$  et  $B$ , qui ensemble couvrent la population d'intérêt  $A \cup B$  comme l'illustre la figure 2. En utilisant la notation de Hartley (1962), il existe trois domaines non chevauchants :  $a = A \cap B^c$ ,  $b = A^c \cap B$  et  $ab = A \cap B$ , où  $c$  désigne le complément d'un ensemble. Les tailles de population des bases de sondage  $A$  et  $B$  sont  $N_A$  et  $N_B$ , et les tailles de population des domaines sont  $N_a$ ,  $N_b$  et  $N_{ab}$ . Nous supposons que  $N_A$  et  $N_B$  sont connues, mais que la taille de population  $N = N_A + N_B - N_{ab}$  peut être inconnue. Dans le présent article, nous supposons que la population, ainsi que les bases de sondage ne varient pas au cours du temps. Il s'agit d'hypothèses fortes, mais dans de nombreuses enquêtes longitudinales, la population d'intérêt et les bases de sondage peuvent être définies pour la période 1. Supposons, à la présente section, que l'appartenance à un domaine est constante au cours du temps. Pour simplifier la notation, nous supposons ici que  $r = 2$  et  $c = 2$ , de sorte qu'il existe deux catégories possibles à chaque période ; le cas général est similaire. Puisque les trois domaines sont non chevauchants, chaque dénombrement de population  $X_{kl}$ ,  $k = 0, 1$ ,  $l = 0, 1$ , peut s'écrire  $X_{kl} = X_{kla} + X_{klb} + X_{klb}$ , où  $X_{klb}$  est le nombre d'unités de la population du domaine  $d$  qui se trouve dans l'état  $k$  à la période 1 et dans l'état  $l$  à la période 2. Les probabilités de population et de domaine correspondantes sont  $p_{kl} = X_{kl}/N$  et  $p_{kla} = X_{kla}/N_A$  pour  $d \in \{a, ab, b\}$ .

Nous tirons des échantillons probabilistes indépendants,  $S_A$  et  $S_B$ , de tailles  $n_A$  et  $n_B$ , des bases de sondage  $A$  et  $B$ . Soit  $w_A^i$  le poids de l'unité échantillonnée  $i$  pour l'échantillon tiré de la base de sondage  $A$  et soit  $w_B^j$  le poids de l'unité échantillonnée  $j$  pour l'échantillon tiré de la base de sondage  $B$ . Nous pouvons donner à  $w_A^i$  la forme d'un poids d'échantillonnage  $[P(i \in S_A)]^{-1}$  ou d'un poids de type Hájek  $[P(i \in S_A)]^{-1} N_A / ($  somme des poids d'échantillonnage dans  $S_A$ ). D'autres scénarios de pondération pour les données longitudinales, discutés dans Verma, Betti et Ghellimi (2007) et dans Lavallée (2007), pourraient également être utilisés. Soit  $y_i = (y_{i1}, y_{i2})$  la réponse de l'unité  $i$  dans  $S_A$ , avec  $y_{i1}, y_{i2} \in \{0, 1, M\}$ , où  $M$  indique que la valeur est manquante. Alors,  $X_A^{kla} = \sum_{i \in S_A} w_A^i I(y_{i1} = k) I(y_{i2} = l) I(i \in a)$  et  $X_A^{klab} = \sum_{i \in S_A} w_A^i I(y_{i1} = k) I(y_{i2} = l) I(i \in ab)$  estiment les dénombrements de population pour la cellule  $(k, l)$  dans les domaines  $a$  et  $ab$  provenant de  $S_A$ , pour  $k, l \in \{0, 1, M\}$ . Soit  $y_j = (y_{j1}, y_{j2})$  la réponse de l'unité  $j$  dans  $S_B$ , et soit

$$X_B^{klb} = \sum_{j \in S_B} w_B^j I(y_{j1} = k) I(y_{j2} = l) I(j \in b) \text{ et } X_B^{klab} = \sum_{j \in S_B} w_B^j I(y_{j1} = k) I(y_{j2} = l) I(j \in ab) \text{ les estimateurs correspondants provenant de } S_B.$$

Dans le présent article, nous supposons que l'appartenance à un domaine peut être déterminée pour chaque unité de l'échantillon et que les réponses  $y_i$  ne contiennent pas d'erreur de classification. Donc, nous supposons que nous savons si chaque unité de l'échantillon tiré de la base de sondage  $A$  ou de la base de sondage  $B$  appartient à l'autre base de sondage ou non. Nous supposons aussi que  $y_i$  et  $y_j$  sont mesurés sans erreur – dans l'exemple de l'emploi, cela signifie que chaque répondant donne la bonne réponse concernant sa situation d'emploi. Donc, les méthodes proposées ici sont sensibles à l'erreur de classification des observations dans les domaines et dans les cellules. Si les moyennes de domaine diffèrent ou si les observations sont classées incorrectement, les estimateurs des flux bruts pourraient présenter un biais ; Pfeffermann et coll. (1998) discutent des méthodes permettant de tenir compte des erreurs de classification dans les enquêtes à base de sondage unique.

Les estimateurs dérivés de  $S_A$  sont présentés au tableau 2. Un tableau similaire peut être conçu pour les estimateurs dérivés de  $S_B$ . Nous supposons que chaque unité est échantillonnée durant l'une des périodes ou les deux. En l'absence de données manquantes, tous les dénombrements estimés pour les cellules  $(k, M)$  et  $(M, l)$  sont nuls. En utilisant l'absence exacte ou approximative de biais des estimateurs, selon que l'on se sert des poids d'échantillonnage ou des poids de Hájek, en l'absence de données manquantes,  $E[X_A^{kla}] \approx X^{kla}$ ,  $E[X_A^{klab}] \approx E[X_B^{klab}] \approx X^{klab}$  et  $E[X_B^{klb}] \approx X^{klb}$ .

Tableau 2  
Estimateurs dérivés de l'échantillon tiré de la base de sondage  $A$

Période 2		Val		Manquante	
Domaine $a$	0	$X_{00a}^A$	$X_{01a}^A$	$X_{01a}^A$	$X_{0Ma}^A$
	1	$X_{10a}^A$	$X_{11a}^A$	$X_{1Ma}^A$	$X_{1+a}^A$
Manquante		$X_{M0a}^A$	$X_{M1a}^A$	$X_{M+a}^A$	
Période 1					
Domaine $ab$	0	$X_{00ab}^A$	$X_{01ab}^A$	$X_{0+ab}^A$	$X_{0Ma+ab}^A$
	1	$X_{10ab}^A$	$X_{11ab}^A$	$X_{1+ab}^A$	$X_{1M+ab}^A$
Manquante		$X_{M0ab}^A$	$X_{M1ab}^A$	$X_{M+ab}^A$	
Val					
3. Estimateurs des flux bruts dans les enquêtes à base de sondage double					
		$X_{+0}^A$	$X_{+1}^A$	$X_{+M}^A$	$\hat{N}_A$

À la présente section, nous dérivons des estimateurs des flux bruts pour des échantillons complexes dans des

Un certain nombre de programmes d'enquêtes longitudinales, telles l'Enquête longitudinale nationale auprès des enfants et des jeunes du Canada et l'Enquête par panel auprès des ménages canadiens, ont maintenant commencé à mettre en œuvre un plan à base de sondage double ou à base de sondage multiple, ou envisagent de le faire. Dans le cas d'une enquête à base de sondage multiple, les échantillons probabilistes sont tirés indépendamment de deux ou plusieurs bases de sondage. L'utilisation de plus d'une base de sondage donne souvent une meilleure couverture de la population et permet de réaliser des économies considérables dans le cas de certaines populations. Ainsi, l'Assets and Health Dynamics Survey (Heerenga 1995), dont l'objectif était d'estimer les caractéristiques de la population de 65 ans et plus, s'appuyait sur un plan à base de sondage double dans lequel la base de sondage *A* était celle d'une enquête nationale auprès de la population générale et la base de sondage *B* était une liste de personnes inscrites au régime Medicare. La structure de cette enquête est illustrée à la figure 1. La base de sondage *A* couvrait l'entière de la population, mais nécessitait une présélection de grande portée afin d'identifier les individus faisant partie de la population cible, ce qui rendait l'échantillonnage coûteux; l'échantillonnage à partir de la base de sondage *B* était moins cher, mais cette base ne contenait pas l'entière de la population. Kalton et Anderson (1986) décrivent l'utilisation de plans à base de sondage double pour échantillonner les populations rares; Blair et Blair (2006) soutiennent que les enquêtes à base de sondage double permettent de tirer parti des modes d'échantillonnage moins coûteux, tels que les méthodes en ligne pour échantillonner les populations rares.

Dans d'autres situations, les deux bases de sondage peuvent être incomplètes, comme l'illustre la figure 2. Hartley (1962, 1974) a été le premier à proposer des estimateurs pour le plan de sondage à base double de la figure 2, quand des échantillons indépendants sont tirés de chaque base de sondage. Des progrès subséquents sont décrits dans Bankier (1986), Fuller et Burmeister (1972), Skinner et Rao (1996), et Lohr et Rao (2000). Lohr et Rao (2006) résument les méthodes d'estimation de quantités de population dans les enquêtes transversales à base de sondage multiple.

Dans le présent article, nous proposons des estimateurs des flux bruts qui peuvent être appliqués aux enquêtes à base de sondage double dans lesquelles l'information longitudinale est recueillie auprès de l'un des échantillons ou des deux. Les unités échantillonnées dans l'une des enquêtes ou dans les deux sont suivies au cours du temps; dans certains cas, des unités supplémentaires sont échantillonnées à des périodes ultérieures afin d'intégrer de nouvelles unités de population ou de compenser l'érosion de l'échantillon. Une enquête longitudinale à base de sondage

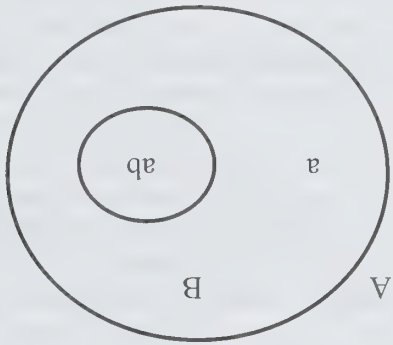


Figure 1 La base de sondage *B* est un sous-ensemble de la base de sondage *A*

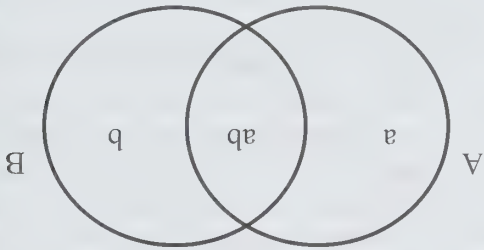


Figure 2 Les bases de sondage *A* et *B* sont toutes deux incomplètes, mais chevauchantes

La présentation de l'article est la suivante. À la section 2, nous exposons le problème de recherche. À la section 3, nous dérivons les estimateurs des flux bruts dans les enquêtes à base de sondage double pour des échantillons complexes pouvant présenter des données manquantes. À la section 4, nous établissons les propriétés asymptotiques et discutons de l'estimation de la variance. À la section 5, nous décrivons une application de notre recherche à la Current



L'estimation des flux bruts dans les enquêtes à base de sondage double

Yan Lu et Sharon Lohr

## Résumé

Les flux bruts sont souvent utilisés pour étudier les transitions concernant la situation d'emploi ou d'autres variables catégoriques chez les individus formant une population. Dans les enquêtes longitudinales à base de sondage double, pour lesquelles des échantillons indépendants sont tirés de deux bases de sondage afin de réduire les coûts d'enquête ou d'améliorer la couverture, l'estimation efficace et cohérente des flux bruts peut poser des défis, à cause des plans de sondage complexes et des données manquantes dans l'un ou l'autre échantillon, ou les deux. Nous proposons des estimateurs des flux bruts dans les enquêtes à base de sondage double et examinons leurs propriétés asymptotiques. Puis, nous estimons les transitions entre les situations d'emploi en utilisant des données provenant de la Current Population Survey et de la Survey of Income and Program Participation.

Mots clés : Enquêtes complexes ; enquêtes à base de sondage double ; jackknife ; estimation longitudinale ; données manquantes.

## 1. Introduction

A l'heure actuelle, de nombreuses enquêtes sont conçues

régulier afin que des quantités longitudinales, telles que les transitions entre les situations d'emploi ou les situations de pauvreté puissent être étudiées. Par exemple, aux États-Unis, la Current Population Survey (CPS : United States Census Bureau 2006) s'appuie sur un plan de sondage à panel rotatif en vertu duquel les personnes habitant l'unité de logement sélectionnée pour l'enquête sont interviewées pendant quatre mois d'affilée, cessent de l'être pendant huit mois, puis sont de nouveau interviewées pendant quatre mois consécutifs. Ce plan permet d'estimer des quantités liées aux changements qui surviennent chez les individus au cours du temps. Puisque de nombreuses réponses aux enquêtes sont catégoriques, les flux bruts, qui sont des transitions entre états d'une variable catégorique au cours du temps, sont particulièrement importants.

Le tableau I donne les nombres d'occurrences d'une variable catégorique mesurée à deux périodes dans une population de  $N$  unités. À la période 1, la variable peut se trouver dans l'un de  $r$  états, et à la période 2, elle peut se trouver dans l'un de  $c$  états. L'exemple qui suit illustre le tableau I. Dans l'étude des changements de situation d'emploi, nous pourrions avoir  $r = 2$  et  $c = 2$ , l'état 0 représentant le chômage et l'état 1, l'emploi. Alors,  $X^{00}$  donne le nombre de membres de la population qui sont en chômage aux deux périodes,  $X^{10}$  est le nombre qui sont occupés à la période 1, mais en chômage à la période 2,  $X^{0+}$  est le nombre total en chômage à la période 1, et ainsi de suite. Nous voulons calculer les estimations et les erreurs-types des flux bruts  $X_{kl}^{0+}$ ,  $k = 0, \dots, r - 1$ ,  $l = 0, \dots, c - 1$ .

Tableau 1  
Table des flux bruts pour la population

Période 2		Période 1	
0	1	2	3
$X_{00}$	$X_{01}$	$X_{02}$	$X_{03}$
$X_{10}$	$X_{11}$	$X_{12}$	$X_{13}$
$X_{20}$	$X_{21}$	$X_{22}$	$X_{23}$
$X_{30}$	$X_{31}$	$X_{32}$	$X_{33}$
$X_{40}$	$X_{41}$	$X_{42}$	$X_{43}$
$X_{50}$	$X_{51}$	$X_{52}$	$X_{53}$
$X_{60}$	$X_{61}$	$X_{62}$	$X_{63}$
$X_{70}$	$X_{71}$	$X_{72}$	$X_{73}$
$X_{80}$	$X_{81}$	$X_{82}$	$X_{83}$
$X_{90}$	$X_{91}$	$X_{92}$	$X_{93}$

a été tiré d'une base de sondage unique.

Même si des estimations transversales successives permettent d'évaluer une variation des taux de chômage au cours du temps, seule une enquête longitudinale permet d'étudier des questions telles que la persistance du chômage chez les individus. L'estimation des flux bruts en utilisant des données d'enquête a été étudiée par de nombreux auteurs, dont Chambers, Woyzbun et Pillig (1988), Hocking et Oxspring (1971), Blumenthal (1968), Chen et Fienberg (1974), Stasny (1984, 1987), ainsi que Stasny et Fienberg (1986). La plupart de ces travaux portaient sur des méthodes en vue d'obtenir des estimateurs du maximum de vraisemblance (EMV) pour les fréquences de cellules attendues dans les tables de contingence en se servant de données partiellement croisées. Pfeiffermann, Skinner et Humphreys (1998) ont proposé des estimateurs qui tiennent compte des erreurs de classification dans les données d'enquête. Tous ces travaux s'appuient sur l'hypothèse qu'un échantillon probabiliste, habituellement un échantillon aléatoire simple,

en utilisant des données d'enquête. En pratique, cet exercice est parfois compliqué à cause des données manquantes et d'autres problèmes.



En vertu du lemme, il faut que  $\bar{\lambda} = 0$ , et  $\bar{M} = \bar{M}$ , donc  $\bar{Q}_{M,\bar{\lambda}} = \bar{Q}_{M'}$ . Il s'ensuit, de nouveau en vertu du lemme, que pour la même constante  $c_n$ , (A.2) est vérifiée si et uniquement si (A.1) est vérifiée pour une valeur donnée de  $\lambda$ . Par conséquent, les modèles compris à l'intérieur de l'enclos, en ce qui concerne  $p$  et  $q$ , sont les mêmes sous les deux méthodes. Il est alors facile de voir, d'après le critère de sélection, que le même modèle  $M_0 = M_0(c_n)$ , en ce qui concerne  $p$  et  $q$ , sera sélectionné sous les deux méthodes pour la constante donnée  $c_n$ . Il s'ensuit alors que la constante  $c_n^*$  sélectionnée en utilisant la méthode adaptative sera la même sous les deux méthodes. Donc, en utilisant de nouveau l'argument susmentionné, le modèle optimal  $M_0^*$ , en ce qui concerne  $p$  et  $q$ , sera le même sous les deux méthodes.

Les formules sous l'expression (7) peuvent être calculées en utilisant les expressions du BLUE et du BLUP (par exemple Jiang 2007, paragraphe 2.3.1) et l'égalité suivante (par exemple, Sen et Srivastava 1990, page 275) : si  $U$  est de dimensions  $n \times q$  et  $V$  est de dimensions  $q \times n$ , alors  $(P + UV)^{-1} = P^{-1} - P^{-1}U(I_q + VP^{-1}U)^{-1}VP^{-1}$  à condition que les inverses existent.

## Bibliographie

- Battese, G.F., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 80, 28-36.
- Chatterjee, S., Lahiri, P. et Li, H. (2007). Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. *Annals of Statistics*, to appear.
- Datta, G.S., et Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.
- Datta, G.S., et Lahiri, P. (2001). Discussions on a paper by Efron & Gous. (Ed., P. Lahiri) *Model Selection*, IMS Lecture Notes Monograph 38.
- Fabrizi, E., et Lahiri, P. (2004). A new approximation to the Bayes information criterion in finite population sampling. Technical Report, Dept. of Math., Univ. of Maryland.
- Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Ganesh, N. (2009). Simultaneous credible intervals for small area estimation problems. *Journal of Multivariate Analysis*, in press.
- Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal (avec discussion). *Statistical Science*, 9, 55-93.
- Hastie, T., et Tibshirani, R.J. (1990). *Generalized Additive Models*. New York : Chapman and Hall.
- Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York : Springer.
- Jiang, J., Rao, J.S., Gu, Z. et Nguyen, T. (2008). Fence methods for mixed model selection. *Annals of Statistics*, 36, 1669-1692.
- Jiang, J., Nguyen, T. et Rao, J.S. (2009). A simplified adaptive fence procedure. *Statistics and Probability Letters*, 79, 625-629.
- Kauermann, G. (2005). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planning and Inference*, 127, 53-69.
- Laird, N.M., et Ware, J.M. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963-974.
- Meza, J., et Lahiri, P. (2005). Une note sur la statistique  $C_p$  sous un modèle de régression à erreur emboîtée. *Techniques d'enquête*, 31, 115-120.
- Morris, C.N., et Christensen, C.T. (1995). Hierarchical models for ranking and for identifying extremes with applications. *Bayes Statistics 5*, Oxford Univ. Press.
- Opsomer, J.D., Breidt, F.J., Claeskens, G., Kauermann, G. et Ranaivosoa, M.G. (2007). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society B*, à paraître.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.
- Ruppert, R., Wand, M. et Carroll, R. (2003). *Semiparametric Regression*. Cambridge Univ. Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Sen, A., et Srivastava, M. (1990). *Regression Analysis*. New York : Springer.
- Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18, 223-249.
- Wang, J.-L. (2005). Nonparametric regression analysis of longitudinal data. *Encyclopedia of Biostatistics*, 2<sup>ème</sup> Ed.

Quelques comparaisons sont toujours utiles. Nous effectuons notre première comparaison avec la méthode de l'enclos proprement dite, mais en utilisant un espace plus restreint de modèles candidats. Plus précisément, nous considérons (12) en imposant la contrainte de splines linéaires uniquement, c'est-à-dire  $p = 1$ , et un nombre de nœuds compris dans l'intervalle de la « règle empirique », c'est-à-dire  $q = 4, 5, 6$ , plus le modèle passant par l'origine ( $p = q = 0$ ) et le modèle linéaire ( $p = 1, q = 0$ ). Dans ces conditions, la méthode de l'enclos donne lieu à la sélection d'une spline linéaire à quatre nœuds (c'est-à-dire  $p = 1, q = 4$ ) comme modèle optimal. La valeur de  $\lambda$  correspondant à ce modèle est approximativement égale à 0,001. Le tracé de  $p^*$  en fonction de  $c_n$  pour cette sélection de modèle est fort semblable au graphique de gauche de la figure 2 et, par conséquent, est omis. En outre, le graphique de droite de la figure 2 donne les valeurs et les courbes prédites sous les deux modèles sélectionnés par la méthode de l'enclos parmi les différents espaces de modèles, ainsi que les points de données originaux.

Une autre comparaison peut être effectuée en traitant (11) comme un modèle additif généralisé (MAG) avec erreurs hétérosédastriques. Nous pouvons obtenir un ajustement pondéré avec degré de lissage optimisé en utilisant un critère de validation croisée généralisée (VCG). Ici, les poids utilisés sont  $w_i = 1/(A + D_i)$ , où l'estimation du maximum de vraisemblance pour  $A$  est utilisée comme estimation par remplacement. Rappelons que les  $D_i$  sont connus. Cette fonction prédite est également superposée dans le graphique de droite de la figure 2. Souignons à quel point elle ressemble à celle prédite par la méthode de l'enclos avec espace restreint.

Pour étendre la classe de modèles pris en considération par lissage fondé sur la VCG, nous avons utilisé la procédure BRUTO (Hasstie et Tibshirani 1990) qui consiste à augmenter la classe de modèles pour examiner un ajustement nul et un ajustement linéaire pour la fonction spline et qui intègre la sélection de modèles résultante (c'est-à-dire modèles nul, linéaire ou lisse) dans un algorithme de rétroajustement pondéré en utilisant le critère VCG pour l'efficacité des calculs. Fait intéressant, ici, la procédure BRUTO trouve simplement un ajustement linéaire global pour la forme fonctionnelle des effets fixes. Il s'agit certes d'une comparaison intéressante, mais les propriétés théoriques de BRUTO pour des modèles tels que (11) n'ont pas vraiment été étudiées en profondeur.

Enfin, comme nous l'avons mentionné à la section 3, en utilisant le lien entre le modèle mixte P-spline et linéaire, nous pouvons formuler (12) sous la forme d'un modèle linéaire mixte, où les coefficients de la fonction spline sont traités comme des effets aléatoires. Le problème se résume alors à un problème de sélection de modèles mixtes

## 6. Conclusion

(paramétriques), de sorte que la méthode de Jiang et coll. (2009) peut être appliquée. En fait, il s'agissait de notre approche initiale pour l'ensemble de données utilisé et le modèle que nous avons trouvé était le même que celui sélectionné par la procédure BRUTO. Cependant, nous avons certaines réserves quant à cette approche, comme nous l'avons expliqué à la section 3.

Bien que le présent article porte principalement sur la sélection de modèles d'EPD non paramétriques, notre méthode pourrait être applicable à des problèmes de sélection de modèles à effets mixtes fondés sur des splines dans d'autres domaines, comme l'analyse de données longitudinales (par exemple Wang 2005).

Dans le cas où un modèle vrai existe parmi les modèles candidats, tels les cas 1 et 2 à la section 4, il est possible d'établir la cohérence de la méthode de l'enclos proposée pour sélectionner le modèle de la même façon qu'à la section 3 de Jiang et coll. (2009) (quoique le résultat de ce dernier article ne s'applique pas directement). Toutefois, en pratique, la situation dans laquelle la modélisation non paramétrique est la plus utile est celle où un modèle vrai n'existe pas, ou qu'il ne figure pas parmi les candidats, comme dans le cas 3 de la section 4. Dans ces conditions, aucun résultat de cohérence ne peut évidemment être prouvé. Il reste à préciser quel serait un comportement asymptotique désirable pour étudier ce dernier cas.

## Remerciements

Les travaux de Jiming Jiang sont financés en partie par les bourses de la NSF DMS-203676 et DMS-0402824. Ceux de J. Sunil Rao sont financés partiellement par les bourses de la NSF DMS-0203724 et DMS-405072, et par la bourse des NIH K25-CA89868.

## Annexe

1. *Preuve du lemme.* Écrivons  $g(\lambda) = \hat{Q}_{M,\lambda}$ . Nous pouvons montrer (détails omis) que  $g'(\lambda) = 2\lambda y' B_\lambda A_\lambda B_\lambda' y$ , où  $A_\lambda = B'(W'W + \lambda BB')^{-1} B$ ,  $B_\lambda = W(W'W + \lambda BB')^{-1} B$  avec  $B' = (0 \ I_q)$  et  $W = (X \ Z)$ . Donc,  $g'(\lambda) \geq 0$  pour  $\lambda > 0$ . En outre,  $\hat{Q}_{M,\lambda} \rightarrow \hat{Q}_M$  quand  $\lambda \rightarrow 0$ .

2. *Preuve du théorème.* Considérons l'inégalité d'enclos

$$\hat{Q}_{M,\lambda} - \hat{Q}_{M,\bar{\lambda}} \leq c_n \quad (A.1)$$

où  $(\bar{M}, \bar{\lambda})$  minimise  $\hat{Q}_{M,\lambda}$ . Considérons aussi l'inégalité d'enclos obtenue en utilisant  $\hat{Q}_M = y' P_{W^*} y$ , qui est

$$\hat{Q}_M - \hat{Q}_{M,\bar{\lambda}} \leq c_n \quad (A.2)$$

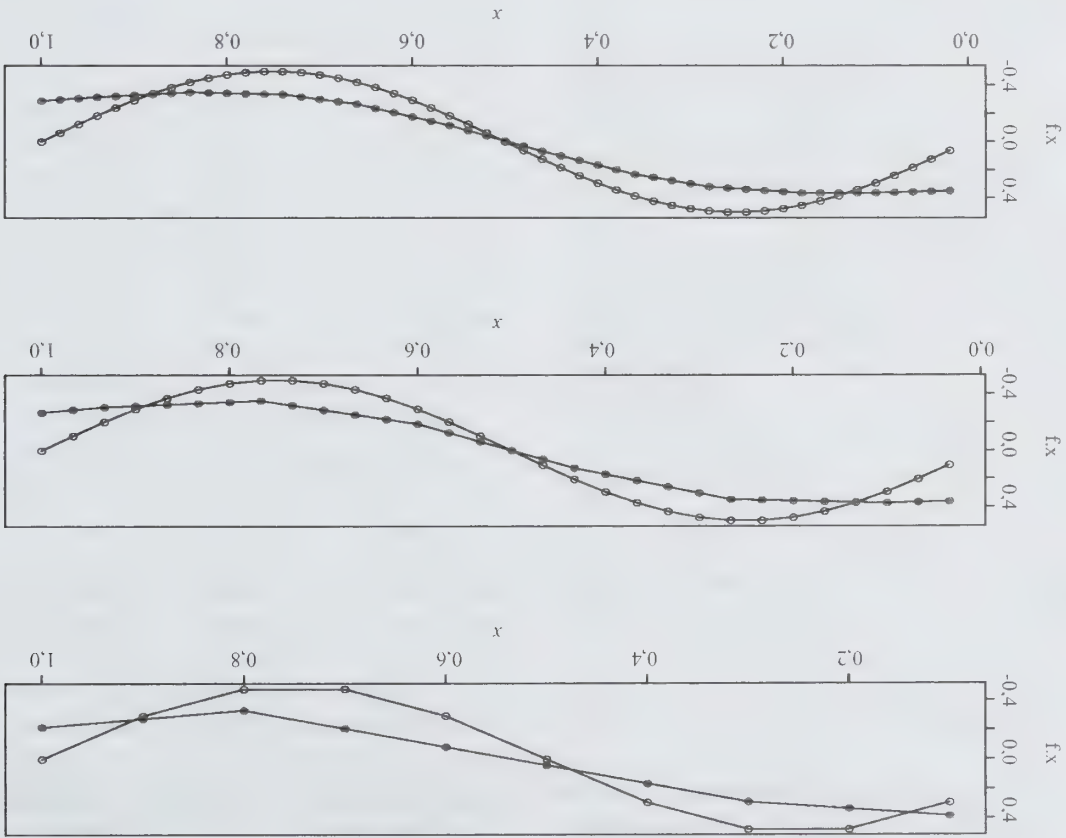


Figure 1 Simulation du cas 3. Graphique du haut : valeurs prédites moyennes pour  $m = 10$ . Graphique du milieu : valeurs prédites moyennes pour  $m = 30$ . Graphique du bas : valeurs prédites moyennes pour  $m = 50$ . Dans tous les cas, les points représentent les valeurs prédites, tandis que les cercles correspondent à la fonction sous-jacente réelle

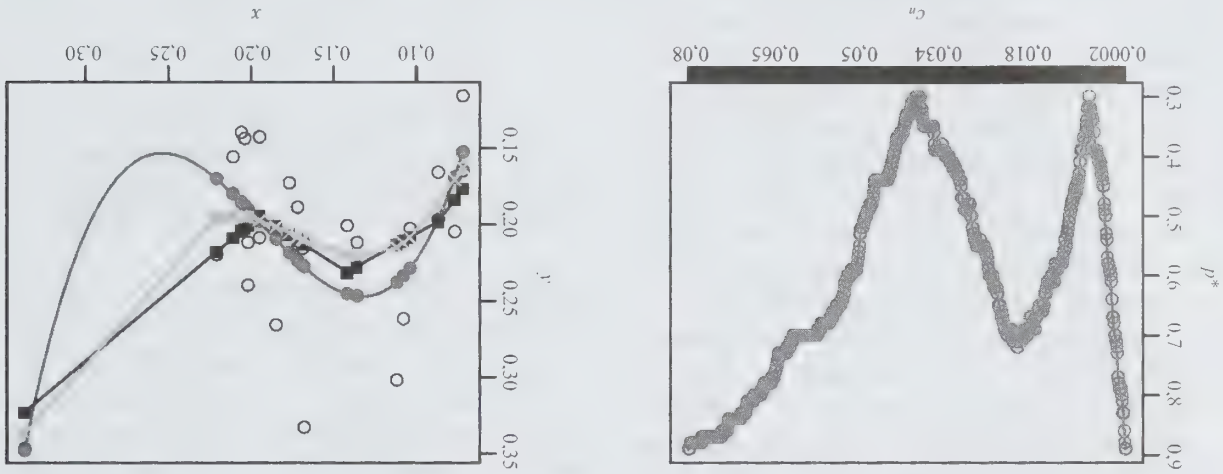


Figure 2 À gauche : graphique de  $p^*$  en fonction de  $p$  pour la recherche sur l'espace complet des modèles. À droite : données brutes et valeurs et courbes prédites ; les points et leur courbe correspondent à la fonction cubique résultant de la recherche dans l'espace complet des modèles ; les carrés et leur courbe correspondent à la spline linéaire avec quatre nœuds résultant de la recherche dans l'espace restreint des modèles ; les  $\circ$  et leur courbe représentent les valeurs prédites par le MAG



Tableau 3  
Données hospitalières tirées de Morris et Christiansen (1995)

Région	$y_i$	$x_i$	$\sqrt{d_i}$
1	0,302	0,112	0,055
2	0,140	0,206	0,053
3	0,203	0,104	0,052
4	0,333	0,168	0,052
5	0,347	0,337	0,047
6	0,216	0,169	0,046
7	0,156	0,211	0,046
8	0,143	0,195	0,046
9	0,220	0,221	0,044
10	0,205	0,077	0,044
11	0,209	0,195	0,042
12	0,266	0,185	0,041
13	0,240	0,202	0,041
14	0,262	0,108	0,036
15	0,144	0,204	0,036
16	0,116	0,072	0,035
17	0,201	0,142	0,033
18	0,212	0,136	0,032
19	0,189	0,172	0,031
20	0,212	0,202	0,029
21	0,166	0,087	0,029
22	0,173	0,177	0,027
23	0,165	0,072	0,025

Ganesh (2009) a proposé pour les taux d'échecs de greffe le modèle de Fay-Herriot suivant :  $y_i = \beta_0 + \beta_1 x_i + v_i + e_i$ , où les  $v_i$  sont les effets aléatoires spécifiques à l'hôpital et les  $e_i$  sont les erreurs d'échantillonnage. Nous supposons que les  $v_i, e_i$  sont indépendants avec  $v_i \sim N(0, A)$  et  $e_i \sim N(0, D_i)$ . Ici, la variance  $A$  est inconnue. En se fondant sur le modèle, Ganesh a obtenu des intervalles crédibles pour les contrastes choisis. Cependant, l'inspection des données brutes suggère certaines tendances non linéaires, de sorte que la question se pose de savoir si l'on peut donner à la partie effets fixes du modèle une forme fonctionnelle plus souple.

Pour répondre à cette question, nous considérons le modèle de Fay-Herriot comme un membre spécial d'une classe de modèles de spline d'approximation dont nous avons discuté à la section 3. Plus précisément, nous supposons que

$$y_i = f(x_i) + v_i + e_i, \quad i = 1, \dots, m, \quad (11)$$

où  $f(x)$  est une fonction lisse inconnue et tous les autres termes sont les mêmes que dans le modèle de Fay-Herriot.

Nous considérons ensuite la classe suivante de modèles de spline d'approximation :

$$\hat{f}(x) = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + \dots + \gamma_q (x - \kappa_q)_+^p \quad (12)$$

avec  $p = 0, 1, 2, 3$  et  $q = 0, 1, \dots, 6$  (le cas  $p = 0$  est choisi selon la « règle empirique » (parce que  $m = 23$ , donc  $m/4 = 5,75$ ). Notons que le modèle de Fay-Herriot correspond au cas où  $p = 1$  et  $q = 0$ . La question est alors de trouver le modèle optimal, en ce qui concerne  $p$  et  $q$ , pour cette classe.

Nous appliquons la méthode de l'enclos adaptatif décrite à la section 3 à ce cas. Ici, pour obtenir les échantillons bootstrap nécessaires pour déterminer  $c_n^*$ , nous commençons par calculer l'estimateur MV sous le modèle  $\tilde{M}$ , qui minimise  $\tilde{Q}_M = y' P_{W \perp V} y$  parmi les modèles candidats [c'est-à-dire (12) ; voir le théorème à la section 3], puis nous tirons des échantillons bootstrap paramétriques sous le modèle  $\tilde{M}$  en traitant les estimateurs MV comme étant les paramètres réels. Cela est raisonnable parce que  $\tilde{M}$  est le meilleur modèle d'approximation pour ce qui est de l'ajustement, même si sous le modèle (11), il pourrait ne pas exister de modèle vrai parmi les modèles candidats. Nous choisissons la taille d'échantillon bootstrap égale à 100.

La méthode de l'enclos donne lieu à la sélection du modèle  $p = 3$  et  $q = 0$ , c'est-à-dire une fonction cubique sans nœuds, comme modèle optimal. Pour être certains que la taille d'échantillon bootstrap  $B = 100$  est adéquate, nous avons répété l'analyse 100 fois, en utilisant chaque fois des échantillons bootstrap différents (rappelons que dans la méthode de l'enclos adaptatif, il faut tirer des échantillons bootstrap afin de déterminer  $c_n^*$ , si bien que la question est de savoir si différents échantillons bootstrap donnent lieu à la sélection de modèles différents). Tous les résultats ont abouti à la sélection du même modèle, à savoir une fonction cubique sans nœuds (même si les quantités intermédiaires dérivées du bootstrap, telles que  $p^*$  et  $c_n^*$ , variaient d'une itération bootstrap à l'autre). Nous avons également effectué l'analyse des données en utilisant  $B = 1\,000$  et le modèle sélectionné est demeuré le même. Donc, il semble que la taille d'échantillon bootstrap de  $B = 100$  est adéquate. Le graphique de gauche de la figure 2 donne le tracé de  $p^*$  en fonction de  $c_n^*$  dans la sélection du modèle par la méthode de l'enclos adaptatif.

différence dans les grands échantillons. Ces constatations corroborent celles de Jiang et coll. (2009).

Tableau 1

Sélection de modèles non paramétriques – cas 1 et cas 2 : probabilités empiriques, en pourcentage, que le modèle optimal soit sélectionné fondé sur 100 simulations

Cas 1				
Taille de l'échantillon	$m = 10$	$m = 15$	$m = 20$	$m = 30$
Pic le plus élevé	62	91	97	83
LI de confiance	73	90	97	96

Cas 2

Taille de l'échantillon	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$
Pic le plus élevé	62	91	97	71	83	97
LI de confiance	73	90	97	73	80	96

Le tableau 2 donne les résultats pour le cas 3. Notons que, contrairement aux cas 1 et 2, il n'existe pas ici de modèle optimal (un modèle optimal doit être un modèle vrai conformément à notre définition). Donc, au lieu de donner les probabilités empiriques de sélection du modèle optimal, nous donnons la distribution empirique des modèles sélectionnés dans chaque cas. Nous voyons que, à mesure que  $\sigma$  augmente, la distribution des modèles sélectionnés est de plus en plus étalée. Nous observons un profil inverse à mesure que  $m$  augmente. La méthode de la limite inférieure (LI) de confiance semble donner de meilleurs résultats pour le choix d'un modèle de spline. Parmi les modèles de spline, la méthode de l'enclos semble donner de façon écrasante la préférence à un petit plutôt qu'à un grand nombre de nœuds.

Soulignons que la méthode de l'enclos nous permet de choisir non seulement  $p$  et  $q$ , mais aussi  $\lambda$  (voir la section 3). Dans chaque simulation, nous calculons  $\hat{\beta} = \hat{\beta}_\lambda$  et  $\hat{\gamma} = \hat{\gamma}_\lambda$ , donné sous l'expression (7), fondé sur le  $\lambda$

Tableau 2  
Sélection de modèles non paramétriques – cas 3 : distributions empiriques, en pourcentage, des modèles sélectionnés

Taille de l'échantillon			Nbre de nœuds					
$m = 10$			0, 2, 3			$m = 30$		
						0, 6, 7, 8		

n'est pas directement nécessaire ici pour le calcul, à cause des deux dernières équations).

Notons qu'à l'étape (ii) du théorème, il n'est pas nécessaire de s'occuper de  $\lambda$ . L'élément qui motive (7) est que cette inégalité est satisfaite quand  $\lambda = 0$ , de sorte que l'on voudrait savoir jusqu'à quel point  $\lambda$  peut aller. En fait, le  $\lambda$  maximal est une solution de l'équation  $\hat{Q}_{M_0, \lambda} - \hat{Q}_M^* = c_n^*$ . Le but des deux dernières équations est d'éviter l'inversion directe de  $V_\lambda = I_n + \lambda^{-1}ZZ'$ , dont la dimension est égale à  $n$ , c'est-à-dire la taille totale de l'échantillon. Notons que  $V_\lambda$  ne possède pas une structure diagonale par bloc à cause de  $ZZ'$ , de sorte que, si  $n$  est grand, l'inversion directe de  $V_\lambda$  peut donner lieu à des calculs difficiles.

La preuve du théorème requiert le lemme qui suit, dont la preuve est donnée en annexe.

**Lemme.** Pour tout  $M$  et  $y$ ,  $\hat{Q}_{M, \lambda}$  est une fonction croissante de  $\lambda$  avec  $\inf_{\lambda > 0} \hat{Q}_{M, \lambda} = \hat{Q}_M^*$ .

#### 4. Simulations

Nous considérons une extension du modèle de Fay-Herriot (Fay et Herriot 1979) dans des conditions non paramétriques. Le modèle peut s'écrire sous la forme

$$y_i = f(x_i) + v_i + e_i, \quad i = 1, \dots, m, \quad (8)$$

où les  $v_i, e_i, i = 1, \dots, m$  sont indépendants et tels que  $v_i \sim N(0, A), e_i \sim N(0, D_i)$ , où  $A$  est inconnue, mais la variance d'échantillonnage  $D_i$  est supposée connue. La principale différence par rapport au modèle de Fay-Herriot classique est  $f(x_i)$ , où  $f(x)$  est une fonction lisse inconnue.

Pour simplifier, nous supposons que  $D_i = D, 1 \leq i \leq m$ . Alors, le modèle peut être exprimé sous la forme

$$y_i = f(x_i) + e_i, \quad i = 1, \dots, m, \quad (9)$$

où  $e_i \sim N(0, \sigma^2)$  avec  $\sigma^2 = A + D$ , qui est inconnue. Donc, le modèle est le même que le modèle de régression non paramétrique.

Nous considérons trois cas distincts qui couvrent divers aspects et situations. Dans le premier cas (cas 1), la fonction sous-jacente réelle est une fonction linéaire,  $f(x) = 1 - x, 0 \leq x \leq 1$ , d'où le modèle se réduit au modèle de Fay-Herriot classique. L'objectif est de découvrir si la méthode de l'enclos permet de valider ce modèle quand il est valide. Dans le second cas (cas 2), la fonction sous-jacente réelle est une spline quadratique à deux nœuds donnée par

$$f(x) = 1 - x + x^2 - 2(x - 1)_+^2 + 2(x - 2)_+^2, \quad 0 \leq x \leq 3 \quad (10)$$

(la forme est un demi-cercle entre 0 et 1 ouvert vers le haut, un demi-cercle entre 1 et 2 ouvert vers le bas et un demi-cercle entre 2 et 3 ouvert vers le haut). Notons que cette

fonction est lisse en ce sens qu'elle possède une dérivée continue. Ici, nous avons l'intention de déterminer si l'enclos permet d'identifier la fonction sous-jacente réelle dans la situation « parfaite », c'est-à-dire quand la fonction  $f(x)$  proprement dite est une spline. Le dernier cas (cas 3), est peut-être celui représentant la situation la plus proche de la pratique, dans laquelle aucune spline ne peut fournir une approximation parfaite de  $f(x)$ . Autrement dit, la fonction sous-jacente réelle ne figure pas parmi les candidates. Dans ce cas,  $f(x)$  est choisie telle que  $0,5 \sin(2\pi x), 0 \leq x \leq 1$ , qui est l'une des fonctions considérées par Kaueermann (2005).

Nous étudions les situations où la taille d'échantillon est petite ou moyenne, à savoir  $m = 10, 15$  ou 20 pour le cas 1,  $m = 30, 40$  ou 50 pour le cas 2, et  $m = 10, 30$  ou 50 pour le cas 3. La covariable  $x_i$  est tirée de la loi uniforme  $[0, 1]$  dans le cas 1, et de la loi uniforme  $[0, 3]$  dans le cas 2, puis maintenue fixe dans toutes les simulations. À l'exemple de Kaueermann (2005), nous posons que  $x_i$  représente les points équidistants dans le cas 3. Nous choisissons pour l'écart-type de l'erreur  $\sigma$  dans (9) la valeur de 0,2 dans les cas 1 et 2. Cette valeur est telle que l'écart-type du signal est, dans chaque cas, à peu près le même que l'écart-type de l'erreur. Pour le cas 3, nous examinons trois valeurs différentes pour  $\sigma$ , à savoir 0,2, 0,5 et 1,0. Ces valeurs sont également du même ordre que l'écart-type du signal dans ce cas.

Les splines d'approximation candidates pour les cas 1 et 2 sont les suivantes :  $p = 0, 1, 2, 3, q = 0$  et  $p = 1, 2, 3, q = 2, 5$  (de sorte qu'il existe, en tout, 10 candidates). Pour le cas 3, comme Kaueermann (2005), nous considérons uniquement des splines linéaires (c'est-à-dire  $p = 1$ ); en outre, nous prenons en considération le nombre de nœuds dans l'intervalle donné par la « règle empirique » (c'est-à-dire environ 4 ou 5 observations par nœud; voir la section 1), ainsi que le modèle passant par l'origine ( $p = q = 0$ ) et le modèle linéaire ( $p = 1, q = 0$ ). Donc, pour  $m = 10$ ,  $q = 0, 2, 3$ , pour  $m = 30, q = 0, 6, 7, 8$  et pour  $m = 50, q = 0, 10, 11, 12, 13$ .

Le tableau 1 donne les résultats fondés sur 100 simulations sous les cas 1 et 2. Comme dans Jiang et coll. (2009), nous considérons à la fois le pic le plus élevé, ce qui consiste à choisir  $c_n^*$  avec  $p^*$  le plus élevé, ainsi que la limite inférieure (LI) de confiance à 95 %, ce qui consiste à choisir une valeur plus faible de  $c_n^*$  correspondant à un pic de  $p^*$  afin d'être prudent, si le  $p^*$  correspondant est supérieur à la limite inférieure de confiance à 95 % de  $p^*$  pour toute valeur de  $c_n^*$  plus grande qui correspond à un pic de  $p^*$ . Nous voyons que la performance de l'enclos adaptatif est satisfaisante même si la taille de l'échantillon est petite. En outre, il semble que la méthode de la limite de confiance inférieure donne de meilleurs résultats dans le cas de petits échantillons, mais ne produit pratiquement pas de



$\gamma$  sans contrainte. Nous avons plutôt  $\hat{Q}_M = |y - X\hat{\beta} - Z\hat{\gamma}|^2$ , où  $\hat{\beta}$  et  $\hat{\gamma}$  sont la solution de (5) et dépendent donc de  $\lambda$ . Le paramètre  $\lambda$  optimal doit être sélectionné par la méthode de l'enclos, en même temps que  $p$  et  $q$ , comme il est décrit plus bas.

Une autre différence tient au fait qu'il pourrait ne pas y avoir de modèle complet parmi les modèles candidats. Par conséquent, nous remplaçons l'inégalité d'enclos (3) par la suivante :

$$(6) \quad \hat{Q}_M - \hat{Q}_M^* \leq c_m,$$

où  $\hat{M}$  est le modèle candidat possédant le  $\hat{Q}_M$  minimal. Nous utilisons le critère d'optimalité à l'intérieur de l'enclos suivant, qui combine la simplicité et le degré de lissage du modèle. Pour les modèles compris dans l'enclos, choisir celui dont la valeur de  $q$  est la plus petite ; s'il existe plus d'un modèle répondant à ce critère, choisir celui dont la valeur de  $p$  est la plus petite. Cela donne le meilleur choix de  $p$  et  $q$ . Une fois que  $p$  et  $q$  sont choisis, nous choisissons le modèle à l'intérieur de l'enclos dont la valeur de  $\lambda$  est la plus grande. De nouveau, notons que  $\lambda$  fait partie du modèle  $M$  qui est sélectionné (ou « estimé ») par la méthode de l'enclos. La constante de réglage  $c_m$  est choisie adaptativement en utilisant la méthode adaptative simplifiée de Jiang et coll. (2009), où le bootstrap par-métrique est employé pour calculer  $p^*$  (voir la section 2).

La preuve du théorème qui suit est donnée en annexe. Par souci de simplicité, supposons que la matrice  $W = (X \ Z)$  est de plein rang. Soit  $P_{W^\perp} = I_n - P_W$ , où  $n = \sum_{i=1}^m n_i$  et  $P_W = W(W'W)^{-1}W'$ .

**Théorème.** En ce qui concerne les calculs, la méthode de l'enclos susmentionnée est équivalente à la procédure suivante : i) d'abord utiliser l'enclos (adaptatif) pour choisir  $p$  et  $q$  en utilisant (6) avec  $\lambda = 0$  et  $\hat{Q}_M = y'P_{W^\perp}y$  (voir le lemme plus bas), ainsi que le critère mentionné plus haut pour choisir  $p, q$  dans l'enclos ; ii) en désignant par  $M_0^*$  le modèle correspondant aux paramètres  $p$  et  $q$  choisis, trouver le  $\lambda$  maximal tel que

$$(7) \quad \hat{Q}_{M_0^*, \lambda} - \hat{Q}_{M_0^*} \leq c_m^*,$$

où, pour tout modèle  $M$  avec les  $X$  et  $Z$  correspondants, nous avons

$$\begin{aligned} \hat{Q}_{M, \lambda} &= |y - X\hat{\beta}_\lambda - Z\hat{\gamma}_\lambda|^2, \\ \hat{\beta}_\lambda &= (X'X - \lambda^{-1}Z'Z)^{-1}X'X^{-1}y, \\ \hat{\gamma}_\lambda &= \lambda^{-1}(I_q + \lambda^{-1}Z'Z)^{-1}Z'(y - X\hat{\beta}_\lambda), \\ X'X^{-1}X &= X'X - \lambda^{-1}X'Z(I_q + \lambda^{-1}Z'Z)^{-1}Z'X, \\ X'X^{-1}y &= X'y - \lambda^{-1}X'Z(I_q + \lambda^{-1}Z'Z)^{-1}Z'y, \end{aligned}$$

et la constante  $c_m^*$  est choisie par la méthode de l'enclos adaptatif décrite à la section 2 ( $c_m$  est définie ci-après, mais

ou  $y = (y_i)_{1 \leq i \leq m}$ , la  $(i, j)^e$  ligne de  $X$  est  $(1, x_{ij}^1, \dots, x_{ij}^p)$ , la  $(i, j)^e$  ligne de  $Z$  est  $[(x_{ij}^1 - \kappa_1^1)', \dots, (x_{ij}^q - \kappa_1^q)']$ ,  $i = 1, \dots, m, j = 1, \dots, n_i$ , et  $\lambda$  est un paramètre de pénalité, ou de lissage. Pour déterminer  $\lambda$ , Wand (2003) a utilisé le lien intéressant suivant avec un modèle mixte linéaire. Pour illustrer l'idée, considérons un cas simple dans lequel  $B_i = 0$  (c'est-à-dire qu'il n'existe pas d'effets aléatoires de petit domaine) et les composantes de  $e_i$  sont indépendantes et suivent une loi  $N(0, \tau^2)$ . Si les  $y$  sont traités comme des effets aléatoires qui sont indépendants et suivent une loi  $N(0, \sigma^2)$ , la solution de (5) est la même que le meilleur estimateur linéaire sans biais (BLUE pour *best linear unbiased estimator*) pour  $\beta$ , et que le meilleur prédicteur linéaire sans biais (BLUP pour *best linear unbiased predictor*) pour  $\gamma$ , si  $\lambda$  est identique au ratio  $\tau^2/\sigma^2$ . Donc, la valeur de  $\lambda$  peut être estimée par les estimateurs du maximum de vraisemblance (MV) ou du maximum de vraisemblance restreint (MVR) de  $\sigma^2$  et  $\tau^2$  (par exemple, Jiang 2007). Cependant, certaines études donnent à penser que cette approche produit un biais de sous-lissage (Kauermann 2005). Considérons, par exemple, le cas particulier dans lequel  $f(x)$  est, en fait, la spline quadratique avec deux nœuds donnés par (10). (Notons que cette fonction est lisse en ce sens qu'elle possède une dérivée continue.) Il est évident que, dans ce cas, la meilleure spline d'approximation devrait être la fonction  $f(x)$  proprement dite avec seulement deux nœuds, c'est-à-dire  $q = 2$  (naturellement, nous pourrions utiliser une spline comportant de nombreux nœuds pour « approximer » la spline quadratique à deux nœuds, mais cela paraît très inefficace dans le cas qui nous occupe). Toutefois, si nous utilisons le lien avec un modèle linéaire mixte mentionné plus haut, l'estimateur MV (ou MVR) de  $\sigma^2$  est convergent uniquement si  $q \rightarrow \infty$  (c'est-à-dire si le nombre d'apparitions des effets aléatoires de la spline tend vers l'infini). L'incohérence apparente à deux conséquences inquiétantes : i) la signification de  $\lambda$  pourrait être conceptuellement difficile à interpréter et ii) le comportement de l'estimateur de  $\lambda$  pourrait être imprévisible.

La méthode de l'enclos offre une approche naturelle pour choisir le degré de la spline,  $p$ , le nombre de nœuds,  $q$ , et le paramètre de lissage,  $\lambda$ , simultanément. Notons toutefois une différence importante par rapport aux situations considérées dans Jiang et coll. (2008) et dans Jiang et coll. (2009) en ce sens que le vrai modèle sous-jacent ne fait pas partie de la classe des modèles candidats, c'est-à-dire les splines d'approximation (1). De surcroît, le rôle de  $\lambda$  dans le modèle devrait être clarifié :  $\lambda$  contrôle le degré de lissage du modèle sous-jacent. Une mesure naturelle du manque d'ajustement est  $\hat{Q}_M = |y - X\hat{\beta} - Z\hat{\gamma}|^2$ . Cependant,  $\hat{Q}_M$  ne s'obtient pas par minimisation de  $\hat{Q}_M$  sur  $\beta$  et

difficile. Parfois, même si l'on peut obtenir une expression pour  $\hat{\Theta}_{M, \tilde{M}}$ , on ne peut garantir son exactitude en tant qu'estimation de l'écart-type dans une situation d'échantillon fini. Jiang, Nguyen et Rao (2009) ont simplifié une méthode de l'enclos adaptative proposée par Jiang et coll. (2008). Pour des raisons de simplicité, nous supposons que  $\mathcal{M}$  contient un modèle complet,  $M_T$ , dont chaque modèle candidat est un sous-modèle. Il s'ensuit que  $\tilde{M} = M_T$ . Dans la méthode adaptative simplifiée, l'inégalité de l'enclos (2)

$$\hat{\Theta}_M - \hat{\Theta}_{M_T} \leq c_n \quad (3)$$

où la constante  $c_n$  est choisie adaptativement comme il suit. Pour chaque  $M \in \mathcal{M}$ , soit  $p^*(M) = P^*\{M_0(c) = M\}$  la probabilité empirique de sélection pour  $M$ , où  $M_0(c)$  désigne le modèle sélectionné par la méthode de l'enclos basée sur (3) avec  $c_n = c$  et  $P^*$  est obtenu par la méthode du bootstrap sous  $M_T$ . Par exemple, sous un modèle paramétrique, on peut estimer les paramètres du modèle sous  $M_T$  puis utiliser un bootstrap paramétrique pour tirer des échantillons sous  $M_T$ . Supposons que  $B$  échantillons sont tirés, alors,  $p^*(M)$  est simplement la proportion d'échantillons (sur un total de  $B$  échantillons) dans lesquels  $M$  est sélectionné par la méthode de l'enclos basée sur (3) avec la constante  $c_n$  donnée. Soit  $p^* = \max_{M \in \mathcal{M}} p^*(M)$ . Notons que  $p^*$  dépend de  $c_n$ . Soit  $c_n^*$  la constante  $c_n$  qui maximise  $p^*$  et qui est celle que nous choisissons. Jiang et coll. (2008) offre l'explication suivante pour justifier la méthode de l'enclos adaptatif. Supposons qu'il existe un modèle vrai parmi les modèles candidats ; le modèle optimal est alors celui à partir duquel les données sont générées et devrait être le plus probable, sachant les données. Donc, sachant  $c_n$ , on recherche le modèle (en utilisant la méthode de l'enclos) qui est le plus appuyé par les données ou, en d'autres termes, celui dont la probabilité (a posteriori) est la plus élevée. Cette dernière est estimée par la méthode du bootstrap. Notons que, bien que les échantillons bootstrap soient générés sous  $M_T$ , ils sont presque identiques à ceux générés sous le modèle optimal. Il en est ainsi parce que les estimations qui correspondent aux paramètres nuls doivent, en principe, être proches de zéro, à condition que les estimateurs des paramètres sous  $M_T$  soient convergents. On dégage alors la constante  $c_n^*$  qui maximise la probabilité (a posteriori) et elle représente le choix optimal.

Il existe deux cas extrêmes correspondant à  $c_n = 0$  et à  $c_n = \infty$  (c'est-à-dire très grand). Notons que, si  $c_n = 0$ ,  $p^* = 1$ , parce que quand  $c_n = 0$ , la méthode aboutit toujours au choix de  $M_T$ . De même, s'il existe un modèle le plus simple unique (par exemple un modèle avec une dimension minimale), disons,  $M_*$ , alors  $p^* = 1$  pour la constante  $c_n$  très grande. En effet, quand  $c_n$  est suffisamment grande, tous les modèles se trouvent dans l'enclos,

Le graphique de droite de la figure 2 en est un exemple. Il s'agit du graphique de  $p^*$  en fonction de  $c_n$  pour l'exemple discuté à la section 5. Ce graphique montre la forme en « W » typique, telle que nous l'avons décrite, et le pic du milieu correspond à l'endroit où se trouve la valeur optimale de  $c_n$ , c'est-à-dire  $c_n^*$ . Jiang et coll. (2009) ont établi la cohérence de l'enclos adaptatif simplifié et étudié ses propriétés en échantillon fini.

### 3. Sélection de modèles d'EPD non paramétriques

Afin de simplifier l'illustration, nous considérons le modèle d'EPD suivant :

$$y_i = f(X_i) + B_i u_i + e_i, \quad i = 1, \dots, m, \quad (4)$$

où  $y_i$  est un vecteur de dimension  $n_i \times 1$  représentant les observations provenant du  $i^{\text{e}}$  petit domaine,  $f(X_i) = [f(x_{ij})]_{1 \leq j \leq n_i}$  avec  $f(x)$  une fonction (lisse) inconnue,  $B_i$  est une matrice connue de dimensions  $n_i \times b$ ,  $u_i$  est un vecteur de dimension  $b \times 1$  d'effets aléatoires propres au petit domaine, et  $e_i$  est un vecteur de dimension  $n_i \times 1$  d'erreurs d'échantillonnage. Nous supposons que les  $u_i$ ,  $e_i$ ,  $i = 1, \dots, m$ , sont indépendants avec  $u_i \sim N(0, G_i)$ ,  $G_i = G_i(\theta)$ , et  $e_i \sim N(0, R_i)$ ,  $R_i = R_i(\theta)$ ,  $\theta$  étant un vecteur inconnu de composantes de variance. Notons que, à part  $f(X_i)$ , le modèle est le même que le modèle mixte linéaire « longitudinal » classique (par exemple Laird et Ware 1982, Datta et Lahiri 2000).

Le modèle de spline d'approximation est donné en remplaçant  $f(x)$  par  $f(x)$  dans (1), où les coefficients  $\beta$  et  $\gamma$  sont estimés par les moindres carrés pénalisés, c'est-à-dire par

$$\text{minimisation de } |y - XB - Z\gamma|^2 + \lambda |\gamma|^2, \quad (5)$$



« règle empirique » générale est que  $p$  est habituellement compris entre 1 et 3, et que  $q$  est proportionnel à la taille d'échantillon,  $n$ , avec 4 ou 5 observations par nœud (Ruppert, Wand et Carroll 2003). Toutefois, étant donné la règle empirique, il pourrait encore demeurer un grand nombre de choix. Par exemple, si  $n = 200$ , les choix possibles pour  $q$  varient de 40 à 50, qui, combinés avec l'intervalle de 1 à 3 pour  $p$ , donnent un total de 33 choix pour la P-spline. Notre nouvelle méthode de l'enclos adaptatif offre une approche dictée par les données pour choisir  $p$  et  $q$  pour le modèle d'EPD fondé sur des splines. La présentation de la suite de l'exposé est la suivante. À la section 2, nous décrivons les méthodes de l'enclos. À la section 3, nous élaborons une méthode adaptative de l'enclos pour résoudre le problème de sélection de modèles non paramétriques. À la section 4, nous démontrons les propriétés en échantillon fini de la nouvelle méthode au moyen d'une série d'études par simulation. À la section 5, nous donnons un exemple fondé sur des données réelles comportant l'ajustement d'un modèle de Fay-Herriot (Fay et Herriot 1979) à un ensemble de données issues d'une enquête médicale. Certains résultats techniques sont présentés en annexe.

## 2. Méthodes de l'enclos

Comme nous l'avons mentionné, le concept fondamental consiste à construire une clôture statistique, puis à sélectionner un modèle optimal parmi ceux se trouvant à l'intérieur de l'enclos en se basant sur un critère d'optimalité, tel que la simplicité du modèle. Soit  $\hat{Q}_M(y; \theta_M)$  représente le vecteur des observations,  $M$  désigne un modèle candidat et  $\theta_M$  désigne le vecteur de paramètres sous le modèle  $M$ . Ici, par manque d'ajustement, nous entendons que  $\hat{Q}_M$  satisfait l'exigence de base que  $E(\hat{Q}_M)$  est minimisée quand  $M$  est un modèle vrai et que  $\theta_M$  est le vecteur de paramètres réels sous  $M$ . Alors, un modèle candidat  $M$  se trouve dans l'enclos si

$$\hat{Q}_M \leq \hat{Q}_M + c_n \hat{\sigma}_{M,M}, \quad (2)$$

où  $\hat{Q}_M = \inf_{\theta_M \in \Theta_M} \hat{Q}_M$ ,  $\Theta_M$  étant l'espace des paramètres sous  $M$ ,  $M$  est un modèle qui minimise  $\hat{Q}_M$  parmi  $M \in \mathcal{M}$ , l'ensemble des modèles candidats, et  $\hat{\sigma}_{M,M}$  est une estimation de l'écart-type de  $\hat{Q}_M - \hat{Q}_M$ . La constante  $c_n$  dans le deuxième membre de (2) peut être choisie comme un nombre fixe (par exemple,  $c_n = 1$ ) ou adaptative-ment (voir plus loin). Le calcul de  $\hat{Q}_M$  est habituellement simple. Par exemple, dans de nombreux cas,  $\hat{Q}_M$  peut être choisi comme la log-vraisemblance négative, ou comme la somme des carrés des résidus. Par ailleurs, le calcul de  $\hat{\sigma}_{M,M}$  peut être assez

Récemment, Jiang et coll. (2008) ont élaboré une nouvelle stratégie de sélection de modèle, qu'ils ont appelée *fence methods*, c'est-à-dire méthodes de l'enclos. Ces auteurs ont constaté que les stratégies classiques de sélection de modèles présentent certaines limites quand elles sont appliquées à des modèles mixtes. Par exemple, la méthode du BIC (Schwarz 1978) s'appuie sur la taille effective d'échantillon dont on n'est pas certain dans des situations typiques d'EPD. Pour illustrer ce point, considérons le modèle de régression à erreurs emboîtées présenté plus haut. Manifestement, la taille effective d'échantillon n'est pas le nombre total d'observations  $n = \sum_{i=1}^m n_i$ , et elle n'est pas proportionnelle non plus à  $m$ , le nombre de petits domaines, à moins que tous les  $n_i$  soient égaux et fixes. Les méthodes de l'enclos permettent d'éviter ces limites et, par conséquent, conviennent pour la résolution des problèmes de sélection des modèles mixtes, y compris les modèles linéaires mixtes et les MLMG. L'idée fondamentale qui sous-tend ces méthodes est de construire une clôture statistique pour isoler un sous-groupe de ce que l'on sait être les modèles corrects. Une fois que la clôture est construite, le modèle optimal est sélectionné parmi ceux se trouvant dans l'enclos en se servant d'un critère qui peut intégrer des quantités d'un intérêt pratiques. Nous donnons ci-après des renseignements plus détaillés sur les méthodes de l'enclos.

Le présent article est axé sur les modèles non paramétriques pour l'estimation sur petits domaines (EPD). Récemment, beaucoup d'attention a été accordée à ces modèles. En particulier, Opsomer, Breidl, Claeskens, Kauermann et Ranaïli (2007) ont proposé un modèle non paramétrique fondé sur des splines pour l'EPD. Leur idée consiste à approximer une fonction moyenne de petit domaine non paramétrique inconnue par une spline pénalisée (P-spline). Ensuite, les auteurs utilisent un lien entre les P-splines et les modèles mixtes linéaires (Wand 2003) pour formuler le modèle d'approximation sous forme d'un modèle linéaire mixte, où les coefficients des splines sont traités comme des effets aléatoires. Pour simplifier, considérons le cas d'une covariable univariée. Nous pouvons alors exprimer une P-spline sous la forme

$$\hat{f}(x) = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \gamma_1 (x - \kappa_1)_p^+ + \dots + \gamma_q (x - \kappa_q)_p^+, \quad (1)$$

où  $p$  est le degré de la spline,  $q$  est le nombre de nœuds,  $\kappa_j$ ,  $1 \leq j \leq q$  sont les nœuds et  $x_+^l = x_l^{(x>0)}$ . Clairement, une P-spline est caractérisée par  $p$ ,  $q$ , ainsi que l'emplacement des nœuds. Notons toutefois que, sachant  $p$  et  $q$ , l'emplacement des nœuds peut être choisi par l'algorithme de remplissage d'espace implémenté dans R [*cover.design*]. La question de savoir comment choisir  $p$  et  $q$  persiste. La



# Méthode de l'enclos pour l'estimation non paramétrique sur petits domaines

Jiming Jiang, Thuan Nguyen et J. Sunil Rao<sup>1</sup>

## Résumé

Nous étudions le problème de la sélection de modèles non paramétriques pour l'estimation sur petits domaines, auquel beaucoup d'attention a été accordée récemment. Nous élaborons une méthode fondée sur le concept de la méthode de l'enclos (*fence method*) de Jiang, Rao, Gu et Nguyen (2008) pour sélectionner la fonction moyenne pour les petits domaines parmi une classe de splines d'approximation. Les études par simulations montrent que la nouvelle méthode donne des résultats impressionnants, même si le nombre de petits domaines est assez faible. Nous appliquons la méthode à un ensemble de données hospitalières sur les échecs de greffe pour choisir un modèle non paramétrique de type Fay-Herriot.

Mots clés : Modèle de Fay-Herriot ; méthode de l'enclos (*fence method*) ; sélection de modèles non paramétriques ; spline pénalisée ; estimation sur petits domaines.

## 1. Introduction

Ces derniers temps, l'estimation sur petits domaines (EPD) fait couler de plus en plus d'encre. Ici, le terme « petit domaine » désigne une population pour laquelle on ne peut produire des statistiques d'intérêt fiables à cause de certaines limites des données disponibles. Une région géographique (par exemple un État, un comté, une municipalité), un groupe démographique (par exemple un groupe particulier âge  $\times$  sexe  $\times$  race) ou un groupe démographique à l'intérieur d'une région géographique sont des exemples de petit domaine. En l'absence d'échantillons directs appropriés pour les petits domaines, des méthodes ont été élaborées afin d'« emprunter de l'information ». Les modèles statistiques, surtout les modèles à effets mixtes, ont joué un rôle important dans l'EPD. Voir Rao (2003) pour un compte rendu complet des diverses méthodes appliquées pour l'EPD.

Alors qu'il existe une abondante littérature sur l'inférence au sujet de petits domaines en utilisant des modèles à effets mixtes, y compris l'estimation des moyennes de petit domaine qui est un problème de prédiction par modèle mixte, l'estimation de l'erreur quadratique moyenne (EQM) du meilleur prédicteur linéaire sans biais empirique (EBLUP pour *empirical best linear unbiased predictor*; voir Rao 2003) et les intervalles de prédiction (par exemple, Chatterjee, Lahiri et Li 2007), beaucoup moins d'attention a été accordée à la sélection du modèle dans le contexte de l'EPD. Pourtant, des chercheurs renommés spécialisés dans ce domaine (par exemple, Battese, Harter et Fuller 1988, Ghosh et Rao 1994) ont souligné l'importance du choix du modèle dans l'EPD. Datta et Lahiri (2001) discutent d'une méthode de sélection de modèles fondée sur le calcul du facteur de Bayes des fréquentistes pour faire le choix entre un modèle à effets fixes

et un modèle à effets aléatoires. Par souci de simplicité, ils se concentrent sur le modèle à effets aléatoires équilibré unidimensionnel suivant :  $y_{ij} = \mu + u_i + e_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, k$ , où les  $u_i$  et les  $e_{ij}$  suivent une loi normale de moyenne nulle et de variance  $\sigma_u^2$  et  $\sigma_e^2$ , respectivement. Comme le soulignent les auteurs, le choix entre un modèle à effets fixes et un modèle à effets aléatoires équilibré, dans ce cas, à tester l'hypothèse unilatérale  $H_0 : \sigma_u^2 = 0$  contre  $H_1 : \sigma_u^2 > 0$ . Toutefois, notons que les problèmes de sélection de modèles ne peuvent pas tous être exprimés sous forme d'un test d'hypothèse. Fahrzi et Lahiri (2004) ont élaboré une méthode robuste de sélection de modèles dans le contexte des enquêtes complexes. Meza et Lahiri (2005) ont démontré les limites de la statistique  $C_p$  de Mallows en sélectionnant les covariables fixes dans un modèle de régression à erreurs emboîtées (Battese, Harter et Fuller 1988) défini par  $y_{ij} = x'_{ij}\beta + u_i + e_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ , où  $y_{ij}$  est l'observation,  $x_{ij}$  est un vecteur de covariables fixes,  $\beta$  est un vecteur de coefficients de régression inconnus, et les  $u_i$  et les  $e_{ij}$  sont les mêmes que dans le modèle mentionné plus haut considéré par Datta et Lahiri (2001). Des études par simulation réalisées par Meza et Lahiri (2005) ont montré que la méthode de la statistique  $C_p$  sans modification ne marche pas bien dans les conditions courantes de modèles mixtes quand la variance  $\sigma_u^2$  est grande ; par ailleurs, un critère  $C_p$  modifié élaboré par ces deux derniers auteurs en ajustant les corrélations intra-grappe donne d'aussi bons résultats que le critère  $C_p$  dans les conditions de régression. Il convient de souligner que toutes ces études sont limitées aux modèles mixtes linéaires, tandis que la sélection du modèle pour l'EPD dans un contexte de modèles linéaires mixtes généralisés (MLMG) n'a jamais été abordée sérieusement.

1. Jiming Jiang, University of California, Davis, Courriel : jiangj@wald.ucdavis.edu ; Thuan Nguyen, (V)regon Health and Science University ; J. Sunil Rao, Case Western Reserve University.



## Prix Cochran-Hansen de 2011

### Concours à l'intention des jeunes statisticiens d'enquête de pays en développement ou en transition

Pour souligner son 25<sup>e</sup> anniversaire, l'Association internationale des statisticiens d'enquête (AISE) a institué le prix Cochran-Hansen, qui est remis tous les deux ans au meilleur article sur les méthodes de recherche par sondage présenté par un jeune statisticien d'un pays en développement ou en transition.

La participation au concours est ouverte aux ressortissants des pays en développement ou en transition qui vivent dans de tels pays et qui sont nés en 1971 ou après.

Les articles soumis doivent être des travaux originaux inédits. Ils peuvent comprendre des parties de la thèse universitaire du participant. Ils doivent être rédigés en français ou en anglais. Les articles présentés doivent être envoyés au Secrétaire de l'AISE à l'adresse ci-dessous et doivent être reçus d'ici le 29 décembre 2010. Chaque article doit être accompagné d'une lettre de présentation qui précise l'année de naissance, la nationalité et le pays de résidence du participant. La lettre de présentation doit également indiquer si l'article soumis est le résultat d'une thèse de doctorat et, dans le cas d'articles communs, le candidat au prix doit indiquer clairement sa contribution à l'article.

Les articles soumis seront examinés par le Comité du prix Cochran-Hansen nommé par l'AISE. La décision du Comité est sans appel.

Le lauréat du prix sera invité à présenter son article à la 58<sup>e</sup> séance de l'Institut international de statistique (IIS), qui aura lieu à Dublin, en Irlande, du 21 au 29 août 2011, et le nom du gagnant sera annoncé lors de l'assemblée générale de l'IIS à Dublin.

L'auteur de l'article gagnant recevra le prix Cochran-Hansen sous forme d'ouvrages et d'abonnements à des périodiques d'une valeur approximative de 500 euros. De plus, des frais de déplacement et de subsistance raisonnables lui seront payés afin qu'il puisse présenter son article à la séance de l'IIS à Dublin.

Pour obtenir de plus amples renseignements, veuillez communiquer avec :

Madame Claude OLIVIER  
Secrétaire de l'AISE  
Association internationale des statisticiens d'enquête  
CEFIL-INSEE, 3 rue de la Cité, 33500 Libourne, France  
Tél. : +33 5 57 55 56 17  
Télec. : +33 5 57 55 56 20  
Courriel : [Claude.olivier@insee.fr](mailto:Claude.olivier@insee.fr)





The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.

# Techniques d'enquête

Une revue éditée par Statistique Canada  
Volume 36, numéro 1, juin 2010

## Table des matières

Prix Cochran-Hansen de 2011 .....	1
<b>Articles Réguliers</b>	
Jimjing Jiang, Thuan Nguyen et J. Sunil Rao Méthode de l'enclos pour l'estimation non paramétrique sur petits domaines .....	3
Yan Lu et Sharon Lohr L'estimation des flux bruts dans les enquêtes à base de sondage double .....	13
Qixuan Chen, Michael R. Elliott et Roderrick J.A. Little Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de population finie dans l'échantillonnage avec probabilités inégales .....	25
David Haziza, Katherine Jenny Thompson et Wesley Yung L'effet des ajustements pour la non-réponse sur l'estimation de la variance .....	39
Jill A. Dever et Richard Valliant Une comparaison des estimateurs de la variance pour la poststratification en fonction de totaux de contrôle estimés .....	49
Patrick J. Farrell et Sarjinder Singh Certaines contributions à la méthode du jackknife appliquée aux estimateurs sous échantillonnage à deux phases .....	63
Jason C. Legg et Cindy L. Yu Comparaison de méthodes de restriction de l'ensemble d'échantillons .....	75
Mojca Bavdaz Modèle multidimensionnel intégral de réponse aux enquêtes auprès des entreprises .....	89
Lazarus Adua et Jeff S. Sharp Examen de la participation aux enquêtes et de la qualité des réponses : l'importance de l'intérêt du sujet et des primes d'incitation .....	103
Tom Krenzke, Lin Li et Keith Rust Evaluation des règles de sélection dans les ménages sous un plan à plusieurs degrés .....	119

*Techniques d'enquête* est répertoriée dans The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods. La revue est également citée par SCOPUS sur les bases de données Elsevier Bibliographic Databases.

## COMITÉ DE DIRECTION

### Président

J. Kovar

### Anciens présidents

D. Royce (2006-2009)

G.J. Brackstone (1986-2005)

R. Plalek (1975-1986)

## COMITÉ DE RÉDACTION

**Rédacteur en chef** M.A. Hidiroglou, *Statistique Canada*

### Rédacteur en

**chef délégué** H. Mantel, *Statistique Canada*

### Rédacteurs associés

J.M. Brick, *Westat Inc.*

P. Cantwell, *U.S. Bureau of the Census*

J.T. Eitinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistique Canada*

D. Judkins, *Westat Inc*

D. Kasprzyk, *Mathematica Policy Research*

P. Kott, *National Agricultural Statistics Service*

P. Lahiri, *JPSM, University of Maryland*

P. Lavallée, *Statistique Canada*

G. Nathan, *Hebrew University*

J. Opsomer, *Colorado State University*

D. Pfeffermann, *Hebrew University*

N.G.N. Prasad, *University of Alberta*

J.N.K. Rao, *Carleton University*

### Rédacteurs adjoints

J.-F. Beaumont, C. Bocci, P. Dick, G. Dubreuil, S. Godbout, D. Haziza, Z. Patak, S. Rubin-Bleuer

et W. Yung, *Statistique Canada*

## POLITIQUE DE RÉDACTION

*Techniques d'enquête* publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

### Présentation de textes pour la revue

*Techniques d'enquête* est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préféablement en Word au rédacteur en chef, (rte@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Turney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca).

### Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada : États-Unis 12 \$ CA (6 \$ × 2 exemplaires); autres pays, 20 \$ CA (10 \$ × 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.

# TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada



# Techniques d'enquête

Une revue  
éditée

par Statistique Canada

Juin 2010 • Volume 36 • Numéro 1

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2010

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 2010

N° 12-001-XPB au catalogue

Périodicité : semestrielle

ISSN 0714-0045

Ottawa



Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à [www.statcan.gc.ca](http://www.statcan.gc.ca). Vous pouvez également communiquer avec nous par courriel à [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca) ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements  
Service national d'appareils de télécommunications pour les malentendants  
Télécopieur 1-800-263-1136  
1-800-363-7629  
Télécopieur 1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements  
Télécopieur 1-613-951-8116  
1-613-951-0581

Programme des services de dépôt

Service de renseignements  
Télécopieur pour le Programme des services de dépôt 1-800-635-7943  
1-800-565-7757

Comment accéder à ce produit ou le commander

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à [www.statcan.gc.ca](http://www.statcan.gc.ca) et de choisir la rubrique « Publications ».

Ce produit n° 12-001-X au catalogue est aussi disponible en version imprimée standard au prix de 30 \$CAN. L'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 \$CAN	12 \$CAN
Autres pays	10 \$CAN	20 \$CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
  - Télécopieur (Canada et États-Unis) 1-877-287-4369
  - Courriel [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca)
  - Poste
- Statistique Canada  
Finances  
Immeuble R.-H.-Coats, 6<sup>e</sup> étage  
150, promenade Tunney's Pasture  
Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site [www.statcan.gc.ca](http://www.statcan.gc.ca) sous « À propos de nous » > « Offrir des services aux Canadiens ».

---

# Techniques d'enquête

---

N° 12-001-XPB au catalogue

Une revue  
éditée  
par Statistique Canada

Juin 2010

•  
Volume 36

•  
Numéro 1





12-001

---

# Survey Methodology

---

Catalogue No. 12-001-XPB

A journal  
published by  
Statistics Canada

December 2010

•

Volume 36

•

Number 2



Statistics  
Canada

Statistique  
Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at [www.statcan.gc.ca](http://www.statcan.gc.ca), e-mail us at [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca), or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

### Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

### Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

## To access and order this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website at [www.statcan.gc.ca](http://www.statcan.gc.ca) and browse by "Key resource" > "Publications."

This product is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca)
- Mail  
Statistics Canada  
Finance  
R.H. Coats Bldg., 6th Floor  
150 Tunney's Pasture Driveway  
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on [www.statcan.gc.ca](http://www.statcan.gc.ca) under "About us" > "The agency" > "Providing services to Canadians."

---

# Survey Methodology

---

A journal  
published by  
Statistics Canada

December 2010 • Volume 36 • Number 2

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2010

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows:  
Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

December 2010

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada  
Statistique Canada

Canada



# SURVEY METHODOLOGY

## A Journal Published by Statistics Canada

*Survey Methodology* is indexed in The ISI Web of knowledge (Web of science). The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

### MANAGEMENT BOARD

**Chairman** J. Kovar

**Past Chairmen** D. Royce (2006-2009)  
G.J. Brackstone (1986-2005)  
R. Platek (1975-1986)

**Members** G. Beaudoin  
S. Fortier (Production Manager)  
J. Gambino  
M.A. Hidirolou  
H. Mantel

### EDITORIAL BOARD

**Editor** M.A. Hidirolou, *Statistics Canada*  
**Deputy Editor** H. Mantel, *Statistics Canada*

**Past Editor** J. Kovar (2006-2009)  
M.P. Singh (1975-2005)

### Associate Editors

J.-F. Beaumont, *Statistics Canada*  
J. van den Brakel, *Statistics Netherlands*  
J.M. Brick, *Westat Inc.*  
P. Cantwell, *U.S. Bureau of the Census*  
R. Chambers, *Centre for Statistical and Survey Methodology*  
J.L. Eltinge, *U.S. Bureau of Labor Statistics*  
W.A. Fuller, *Iowa State University*  
J. Gambino, *Statistics Canada*  
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*  
D. Judkins, *Westat Inc.*  
D. Kasprzyk, *Mathematica Policy Research*  
P. Kott, *National Agricultural Statistics Service*  
P. Lahiri, *JPSM, University of Maryland*  
P. Lavallée, *Statistics Canada*  
P. Lynn, *University of Essex*  
D.J. Malec, *U.S. Census Bureau*  
G. Nathan, *Hebrew University*  
J. Opsomer, *Colorado State University*

D. Pfeffermann, *Hebrew University*  
N.G.N. Prasad, *University of Alberta*  
J.N.K. Rao, *Carleton University*  
J. Reiter, *Duke University*  
L.-P. Rivest, *Université Laval*  
N. Schenker, *National Center for Health Statistics*  
F.J. Scheuren, *National Opinion Research Center*  
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*  
P. Smith, *Office for National Statistics*  
E. Stasny, *Ohio State University*  
D. Steel, *University of Wollongong*  
L. Stokes, *Southern Methodist University*  
M. Thompson, *University of Waterloo*  
V.J. Verma, *Università degli Studi di Siena*  
K.M. Wolter, *Iowa State University*  
C. Wu, *University of Waterloo*  
W. Yung, *Statistics Canada*  
A. Zaslavsky, *Harvard University*

**Assistant Editors** C. Bocci, P. Dick, G. Dubreuil, S. Godbout, D. Haziza, Z. Patak and S. Rubin-Bleuer, *Statistics Canada*

### EDITORIAL POLICY

*Survey Methodology* publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

### Submission of Manuscripts

*Survey Methodology* is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca. Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site ([www.statcan.gc.ca](http://www.statcan.gc.ca)).

### Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: [www.statcan.gc.ca](http://www.statcan.gc.ca).

**Survey Methodology**  
A Journal Published by Statistics Canada  
Volume 36, Number 2, December 2010

**Contents**

**Waksberg Invited Paper Series**

Ivan P. Fellegi

The organisation of statistical methodology and methodological research in national statistical offices ..... 123

**Regular Papers**

Carl-Erik Särndal and Sixten Lundström

Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias ..... 131

Jae Kwang Kim

Calibration estimation using exponential tilting in sample surveys ..... 145

Stephen J. Haslett, Marissa C. Isidro and Geoffrey Jones

Comparison of survey regression techniques in the context of small area estimation of poverty ..... 157

Maria Rosaria Ferrante and Carlo Trivisano

Small area estimation of the number of firms' recruits by using multivariate models for count data ..... 171

Julia D'Arrigo and Chris Skinner

Linearization variance estimation for generalized raking estimators in the presence of nonresponse ..... 181

Abdellatif Demnati and J.N.K. Rao

Linearization variance estimators for model parameters from complex survey data ..... 193

Kirk M. Wolter, Phil Smith and Stephen J. Blumberg

Statistical foundations of cell-phone surveys ..... 203

**Short Notes**

Rudolf Witt, Diemuth E. Pemsil and Hermann Waibel

Collecting data for poverty and vulnerability assessment in remote areas in Sub-Saharan Africa ..... 217

Mohamed G. Qayad, Pranesh Chowdhury, Shaohua Hu and Lina Balluz

Respondent differences and length of data collection in the Behavioral Risk Factor Surveillance System ..... 223

Yves Tillé and David Haziza

An interesting property of the entropy of some sampling designs ..... 229

**Acknowledgements** ..... 233

**Announcements** ..... 235

**In Other Journals** ..... 237

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.





## Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work.

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2012 Waksberg Award.

This issue of *Survey Methodology* opens with the tenth paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Leyla Mohadjer (Chair), Daniel Kasprzyk, Elisabeth A. Martin and Wayne Fuller for having selected Ivan P. Fellegi as the author of this year's Waksberg Award paper.

### 2010 Waksberg Invited Paper

**Author: Ivan P. Fellegi**

Ivan P. Fellegi is Chief Statistician of Canada Emeritus at Statistics Canada. He was the Chief Statistician of Canada from 1985 to 2008, and it was during that period that Statistics Canada was ranked by *The Economist* as the best statistical office in the world. Dr. Fellegi contributed significantly both to survey methodology and to the effective management of a large organization during his long career at Statistics Canada.

He has published extensively on statistical methods, on the social and economic applications of statistics and on the successful management of statistical agencies. Some of his methodology papers have become landmarks: topics covered include sample design, edit and imputation, record linkage, and the analysis of survey data. He has actively participated on several committees: he was chair, Conference of European Statisticians of the United Nations Economic Commission for Europe (1993-97); Chair of the Committee on Statistics of the Organisation for Economic Cooperation and Development (2004-2008); past President of the International Statistical Institute, the International Association of Survey Statisticians, and the Statistical Society of Canada; and past chair of the Board of Governors, Carleton University (1995-97). He has a long list of honours that include: Officer of the Order of Canada; recipient of the Outstanding Achievement Award of the Public Service of Canada; the Order of Merit of the Hungarian Republic; the Career Achievement Award of the Canadian Policy Research Initiative, La Médaille de la ville de Paris, Member of the Hungarian Academy of Sciences, Gold Medal of the Statistical Society of Canada and the Robert Schuman medal of the European Community. He is the recipient of Honorary Doctorates from Université de Montréal, Université du Québec (Institut national de la recherche scientifique), Simon Fraser University, McMaster University, Carleton University, and the University of Ottawa. He is an Honorary Member of the International Statistical Institute, Honorary Fellow of the Royal Statistical Society.



# The organisation of statistical methodology and methodological research in national statistical offices

Ivan P. Fellegi<sup>1</sup>



## Abstract

The paper explores and assesses the approaches used by statistical offices to ensure effective methodological input into their statistical practice. The tension between independence and relevance is a common theme: generally, methodologists have to work closely with the rest of the statistical organisation for their work to be relevant; but they also need to have a degree of independence to question the use of existing methods and to lead the introduction of new ones where needed. And, of course, there is a need for an effective research program which, on the one hand, has a degree of independence needed by any research program, but which, on the other hand, is sufficiently connected so that its work is both motivated by and feeds back into the daily work of the statistical office. The paper explores alternative modalities of organisation; leadership; planning and funding; the role of project teams; career development; external advisory committees; interaction with the academic community; and research.

Key Words: Methodology; Official statistics; Statistical organisation; Research; Relevance; Independence.

## 1. Introduction

It is a great honour to accept an award named after Joe Waksberg. Joe has been a close personal friend, as well a good friend of Statistics Canada.

I came to know Joe during his latter years in the Bureau of the Census when Morris Hansen asked me to become a member of what was then a most imposing methodology advisory committee of the Bureau chaired by Bill Cochran. Subsequently, in the late 1970s, when Statistics Canada had serious problems of image and of internal management, Statistics Canada asked a group of prominent statisticians to review what was wrong. At my recommendation, Joe was one of the three wise men asked to take part (the others being Richard Ruggles and the chairman, Claus Moser). Joe immediately agreed and in his inimitable low-key manner made invaluable contributions to Statistics Canada; the very helpful basic message being that while we had serious management problems, there was nothing much wrong with our methodology.

A few years ago the Census Bureau honoured me by asking to give one of their annual “wise elders” lectures. While I objected strongly on the grounds that I neither considered myself “wise”, nor “elder”, in the end I accepted their kind invitation. With typical grace, Joe took the time to show up for my talk, even though he was well into the middle of his eighties but still very busy as chairman of the board of WESTAT. We had a really good chat; and that was the last time I saw him. What a career; what a life!

So it is not only a professional honour to accept the Waksberg Award, but also a personal pleasure to be associated with Joe one more time.

I was told that generally the recipients of the Waksberg Award give an overview of an area of methodology. But while, as you know, I did spend the first half of my career as a methodologist, I stopped being a practitioner some decades ago – although I am still an ardent advocate (see Fellegi 2004). So I thought I would join the first half of my career – methodology – to the second half – management of statistical offices. I shall therefore, talk about the lessons I learnt about the organisation of applied methodological work and methodology research in national statistical offices; what works well and what less so (I assume that the basic conditions for an effective methodology function exist: there is a supply of trained statisticians in the country, the statistical office has a functioning infrastructure, salaries, if they are not competitive, are at least within sight of what is offered in the private sector, and so on).

I have two overall themes. Managing the tension between independence and relevance is one of them: generally, methodologists must work closely with the rest of the statistical organisation for their work to be relevant. Indeed, they must strive to serve the objectives of external clients, represented inside the office by subject matter experts. However, for them to be effective they must enjoy the necessary independence to question the use of existing methods, and to champion new ones if they believe they could reduce costs or increase statistical quality.

But the effectiveness of methodology also depends on a strong methodology research capacity which, on the one hand, has the necessary independence needed by any research program, but which, on the other hand, is sufficiently connected to on-going work so that it is both motivated by and feeds back into the daily practice of the

1. Ivan P. Fellegi is Chief Statistician Emeritus at Statistics Canada. E-mail: [ivan.p.fellegi@statcan.gc.ca](mailto:ivan.p.fellegi@statcan.gc.ca).



statistical office. The organisation of methodology research will be my second them.

But first I want to define what I mean in the present context by the terms *methodology*, *relevance* and *independence*.

## 2. Some definitions

### *Methodology*

The unique service performed by *methodology* is to maximise statistical quality given an imposed budget (or conversely). They do so through the application of statistical practice that is either based on statistical theory or on organized empirical observation. In other words methodologists are wizards of the relevant statistical theories; but also of “organised empirical observation” where formal theory abandons us. By organised empirical evidence I mean designed experiments or analytically assessed experience. So I am including all organized knowledge about the use of methods and approaches that result in the objective of maximising quality within a budget – or conversely, minimising the budget needed to achieve a stated quality level.

This would include such things as sample design, estimation, data editing, imputation, exploitation of administrative data, record linkage, seasonal adjustment, questionnaire design, measurement of accuracy and quality assurance of censuses and surveys, the use of experimental designs, and so on.

Methodologists are predominantly mathematical statisticians and they work on the applied end of their subject. Because of the interdisciplinary nature of official statistics they interact with survey managers, experts in data collection, IT personnel, geographers, sociologists, economists, etc.

### *Relevance*

Methodology is *relevant* if the day to day practice of the statistical office is actually based on sound methodology. A major issue in the organization of methodology is how to balance the intrinsically service nature of methodology against the need for the function to provide strong and effective guidance. Much of the paper will deal with all those arrangements needed to ensure the objective of relevance.

In the case of methodological *research*, relevance means that the research is both motivated by and informs applied work.

### *Independence*

The notion of independence of methodology means the ability to provide sound methodological guidance to projects, irrespective of the hierarchical arrangement of line

organisations that *can be debated but not ignored*; and that this debate is based on evidence, not authority. So my definition of independence is not that methodologists should be able to “do their own thing” but rather that they should have an authoritative voice.

Independence is frequently contrasted with relevance. Since relevance is about embedding methodology into practice, this is often attempted by building methodological services right into the fabric of subject matter organisations. By contrast, independence is thought to be enhanced by giving methodologists their own organisation(s). In this sense, therefore, there is a tension between the two. However, I would argue that relevance cannot be achieved if methodological guidance is ignored, so appropriate arrangements to ensure independence are necessary for relevance.

Independence of *methodological research* is different: it is generally meant to refer to an environment in which researchers have predominant say in the choice of their topics. Clearly, providing researchers with such an environment does create a permanent tension with the need to be relevant at all times, particularly when it is not at all obvious in the short term where the relevance lies.

In my discussion of how to balance relevance and independence of both the applied methodology function and of methodology research I will describe not only organisational arrangements, but a wide variety of tools and arrangements that should be considered in the pursuit of this objective. I shall use Statistics Canada as a concrete illustration. What I wish to emphasize is that the issue is much more complicated than what the terms “centralisation” and “decentralisation” denote for whichever of these basic organisational arrangements is adopted, many additional tools are needed to offset their disadvantages while maintaining their intrinsic advantages. Indeed, I have organised the rest of the paper around a discussion of the main tools (in choosing these tools for discussion, I borrowed from the paper by Brackstone 1997) involved under the following headings:

- Organisation;
- Leadership;
- Planning and funding;
- Project teams;
- Career development;
- Advisory Committees;
- Interaction with the academic community; and
- Research.

## 3. Organisation

### *General thoughts*

National statistical offices differ in the way they organise their methodology functions. In some it is distributed to

individual parts of the agency, each responsible for a given subject (e.g., labour). In other agencies decentralisation is only partial, e.g., to broader subject matter areas (such as demography or business statistics). The US Bureau of the Census, for example has largely decentralised its methodology function. By contrast, Statistics Canada and the Australian Bureau of Statistics have largely centralised it. Many factors influence the organizational choice. For example, in France and in India where all professionals share similar background in statistics and are largely recruited from a single teaching institution the accent is obviously on centralizing training and to a lesser extent research.

The traditional arguments are that decentralisation favours relevance and centralisation favours independence. However, the aim should be to have both. That being the case the question is how we can enhance independence in the case of decentralised methodology organisations, and relevance in the case of centralised ones.

Decentralisation, while potentially serving to underscore relevance, has some built-in disadvantages. Since each unit to which methodology is decentralized is necessarily smaller than it would be in more centralized options, it is less likely to facilitate specialisation and research. It is also less likely to encourage cross-fertilisation by methodologists working on other issues. Also, since the line organisations to which methodology is decentralised are typically not headed by methodologists, this model tends to result in lower hierarchical positions for the heads of the decentralised methodology units. In case of “conflicts” – and these will be inevitable because of different perceptions of priority, cost, quality and so on – other things being equal it will be more difficult for methodologists to defend their professional advice. If left without a counterweight, this kind of organization could get out of balance.

A critical counterweight could be a “chief methodologist” who reports directly to the head of the statistical office and inevitably is called upon to play an important role in long term planning and resource allocation. The “Chief Methodologist” could have his hand strengthened if given direct line responsibility for a strong research and development function which could serve as the “intellectual home base” for the decentralised methodology staff.

Project teams, brought together for large developments, are another important tool to enhance independence in the case of centralised organisations. Such projects – which if at all significant are necessarily multi-disciplinary – are carried out by ad hoc project teams which operate off-line from the agency’s line organization. The organization of project teams is a matter to which Statistics Canada devoted considerable attention and it has been refined over time. Among its elements there is the feature that whenever

professional disputes within the teams arise and the team believes that their solution requires outside intervention, the dispute is referred to a senior group of which someone from the staff of the “chief methodologist” is a member (this is automatically the case if the methodologist comes from a centralised group). It is this senior steering group that can contribute to protecting independence.

Consideration might also be given to providing some additional tools for the “chief methodologist”: he could be authorised and funded to develop a strong methodology training program; he could be given a strong role in the allocation and career development of the methodology staff; he could be supported by a strong external advisory committee; and so on. These features recognize that the role of “chief methodologist” is particularly delicate and could become more so if his place in the hierarchy were dependent on the size of the staff he controls directly without provision – as there is in some countries – to have his level of access and place in the ladder depend on his personal prestige rather than on the size or level of supporting staff.

#### *Centralisation: the Statistics Canada model*

Many years ago Statistics Canada opted for the centralised model (see Fellegi 1996) and that option was never seriously challenged (it was challenged for a brief period of time in the late seventies but in concrete terms the challenge did not get anywhere), and put in place a number of practices designed to reduce the threat that centralisation might result in diminished relevance.

1. Project teams: These are inter-disciplinary and include as a matter of course a methodologist but they are headed by a project manager whose association with the project is subject matter and who is likely to assume operational responsibility for the completed project.
2. Funding: much of the funding for the methodology function is controlled by the rest of Statistics Canada. Program areas (within limits that I will describe further) are free to spend their money on buying methodology services or not so long as they do not fall foul of the agency’s quality norms and accepted standards. With their budget largely on the line year after year, this accountability means that it is very much in the interest of methodologists to be responsive to the needs of the Agency’s Programs.
3. Organisation of the methodology function: it largely parallels the organisation of Statistics Canada. There are four methodology divisions: three of them provide methodology input to three different areas of the agency, while the fourth is devoted to research. In



fact, the three applied methodology divisions are themselves organised by subject matter in parallel with the manner in which the bureau is organised (regular rotation of methodology staff ensures broad development opportunities for methodologists).

4. Co-location of methodology staff: methodologists are occasionally physically moved to the offices of the subject matter areas whose surveys they help to design. This is an additional measure taken to ensure that they focus on the right issues.
5. Finally, as a matter of sound practice, methodologists conduct – and follow up on the results of – client satisfaction surveys which provide feedback on all aspects of their performance and first and foremost on the relevance thereof.

#### 4. Leadership

##### *General thoughts*

Leadership is crucial. The leader of the methodology function, in addition to a proper academic background and a great deal of experience in methodology, must possess a strategic vision and a personality that inspires confidence. This is an intrinsically difficult function. In the overwhelming majority of offices operational and subject matter considerations are the ones that receive the most attention. In such an environment an authoritative voice for methodology is needed to ensure adequate resources for the methodology function itself, but even more importantly to lead the *entire agency* in directions that are technically sound, and conversely to hold back initiatives that cannot be supported by sound methodology. “Soundly based” involves more than good survey design that uses the best available current knowledge. It also includes the notion of strategic planning of research, experiments and pilot surveys so as to improve the likelihood that whatever knowledge will be needed in the future will be available. For the opinions of methodologists to make a proper impact they must be supported by a leader whose unchallenged personal competence is combined with a seat at the statistical agency’s most senior table.

If methodologists do not belong to a central organization within the statistical agency it is all the more important for their senior representative to be highly placed in the hierarchy since under a decentralized scheme he would not have direct line authority for (the bulk of) methodology resources.

##### *Centralisation: the Statistics Canada model*

Centralisation provides another lever to enable the leader of the methodology function to carry out his proper role as it

enables him to make rational and authoritative assignments of the resources under his direction to the most strategic projects. The top advocate of sound methodology in Statistics Canada has the status of Assistant Chief Statistician (ACS) – the rank immediately below that of the Chief Statistician of Canada. In order to secure such a high position in a government bureaucracy, the line responsibility of the ACS (Methodology) includes statistical standards (classifications and central registers), as well as informatics (IT). While the position is therefore responsible for more than methodology, it is by long tradition (over 35 years) filled by someone who is a noted expert on methodology and can therefore speak at the top table authoritatively about its importance in general as well as in the context of particular projects.

#### 5. Planning and funding

##### *General thoughts*

The effective functioning of methodology (as indeed the entire statistical office) greatly depends on the existence of a proper planning system (see Fellegi 1992 and Brackstone 1991):

- Planning is a necessary condition to ensure that resources are allocated rationally at all times.
- It also serves to mark explicitly the beginning and the end of development projects and therefore constitutes the ideal opportunity for methodology to “sign off” on the proposed design of new projects.
- Lastly, the planning system creates an opportunity for methodology to make an explicit judgement on whether a planned new venture can respect simultaneously its budgetary constraints, the agency’s quality standards, and the expected maintenance bill. In fact, the planning system also provides an opportunity for all representatives of the disciplines involved in the creation of a new project (its planning or its implementation) to “sign off” as a mark of assuming professional responsibility for the adequacy of its funding or for the integrity of its functioning.

Such a planning system is essential where the main disciplines (methodology, systems development, data collection, *etc.*) are centralised for otherwise the organisations responsible cannot make provisions for the needed resources. But, for more subtle reasons, decentralised offices need it just as much: to provide an explicit forum for the leaders of methodology (and, indeed, other key disciplines), to make their input during the critical formative stages of new projects.



### *Centralisation: the Statistics Canada model*

Every new project or major redesign is approved within Statistics Canada's planning system. In preparation for its consideration, a comprehensive budget is developed and all major disciplines which are required to contribute sign off on the appropriateness of the proposed design and operational modalities. If the project is approved, its budget is divided up and distributed to participating disciplines, including methodology. In turn, these organisations "contract" to deliver the agreed contributions within the approved budgets. A project manager oversees both progress and expenditures, with authority to reassign resources, if necessary.

The budget of the Methodology organization is composed of five distinct sources. These are designed, on the one hand, to facilitate the sound planning of the use of methodology and its thorough integration into the work of the Agency, and on the other to secure for it the needed funding.

1. The contribution of methodology to *developmental* projects is guaranteed by the planning process of Statistics Canada, as indicated above. The financial contribution to the methodology budget from these sources may vary from year to year, but there is a reasonable overall stability (facilitating the hiring and development of permanent staff). They account for almost 30 per cent of the total methodology budget. These projects typically involve major redesigns, often with significant experimentation and innovation.
2. But methodological contributions are also needed for maintenance (quality control, monitoring of various errors including variance estimation where relevant, minor design adjustments, *etc.*). For these activities there are core resources set aside and more or less permanently allocated by broad subject matter. This constitutes the second component of the methodology budget and it accounts for somewhat less than 25%. While for methodology this "on-going" work accounts for less than 25% of their workload, for Statistics Canada as a whole "on-going" work accounts for over 90% of our budget. This is because of the innovative nature of methodology work.
3. A third component comes from supplementary resources funded directly by the beneficiary subject matter divisions who, in effect, make savings from their other expenditures to avail themselves of additional methodology contributions. These supplementary funds account for a by no means negligible 20% or so of the methodology budget. The very fact that subject matter divisions consider

methodology sufficiently valuable to fund methodological advice directly says a lot about the health of the relationship and of the extent to which it is valued. The funds in question are for a mixture of projects including enhancements short of a major redesign of on-going projects. They also strengthen the awareness of methodology staff of the need to remain relevant for their users. The kind of service they provide has a direct bearing on the amount of resources that are made available to them.

4. The fourth part of the methodology budget (about 20 per cent) comes from externally funded projects, typically from the budget of surveys funded by other departments. No more needs to be said about them.
5. The final part (7 per cent) is for research. This is a "block fund", meaning that a certain fixed amount of funds is allocated for the research function. The annual allocation is governed by a mechanism described below.

The intricacies of the funding mechanism and the multiplicity of funding sources are a reflection of the care exercised in the agency to balance the virtues of independence with those of relevance.

## **6. Project teams**

### *General thoughts*

The use of project teams in developmental projects helps to strengthen relevance without it being necessarily at the expense of independence. But project teams are not a universal panacea as everything depends on establishing appropriate checks and balances. In centralised organisations project teams, most often headed by a project manager from the sponsoring subject matter area, help to nudge the participating methodology staff to pay proper attention to the objectives and constraints of projects. Nonetheless there remains an inherent danger that the project manager will not give sufficient weight to the considered advice of methodologists.

Project teams in decentralised organisations are just as important to ensure that the views of methodologists are given appropriate weight. Here, however, the dice are clearly weighted in favour of relevance and against independence. Moreover, an exaggerated emphasis on "relevance" has its danger as well since it can lead to local optimisation. Local optimisation is a situation where surveys are optimised without regard to agency wide objectives. An example might be a situation where surveys are customised to an extent such that the introduction of important efficiencies through the use of agency-wide

standards and general systems becomes difficult (the widespread use of generalized approaches, systems and tools can be a source of considerable agency-wide efficiencies: they shorten implementation times, reduce the expenditure on both systems development and maintenance, facilitate staff rotation, *etc.* However, generalized systems might lack some features which could enhance the efficiency of any given operation. Decentralized organizations are more likely to favour such locally developed solutions in preference to agency-wide standard tools, even though the latter might lead to substantial *long-run* efficiencies).

#### *Centralisation: the case of Statistics Canada*

In Statistics Canada project teams working on major development projects are accountable and report to steering committees typically composed of the heads of the participating disciplines. A steering committee approves the broad project strategy, and serves, if needed, as a forum to which issues can be referred that could not be resolved within the team itself. In practice such appeals are rare and are restricted to cases where professional principles or truly strategic issues are involved. Steering committees ensure that issues do not get resolved within the project team on the basis of rank but rather on the basis of professional merit.

Methodologists serving on project teams carry out a dual function:

- At a strategic level, they help ensure that the overall survey design achieves the project's substantive objectives, while striking a balance between reliability, cost, timeliness and respondent burden. While striving for this balance concerns the entire project team it is the methodologists who provide the framework and techniques that must be considered in seeking the optimum balance.
- At a tactical level the methodologists provide the statistical methods and tools that are incorporated into the overall survey design: the sample design, the estimation and weighting approach, quality control, editing and imputation strategies, coverage checks, analytic methods and the like.

Project teams function best in an organisation dedicated to making decisions on the basis of merit; where everyone can pose questions and expect reasoned answers; one that is devoted to making maximum use of the expertise of everyone involved.

## **7. Career development**

### *General considerations*

Career development is essential for all professional groups, and it involves both formal training as well as

formal and informal approaches to facilitate on-the-job learning. Methodology staff, in my view, requires special attention in this respect. The reason is that universities in general offer few, if any, courses in survey methodology (there is an increasing number of exceptions, although their numbers are still far from overwhelming. A most notable one is the Joint Program in Survey Methodology, University of Maryland. But there are also degree programs on official statistics in the UK, Ireland and New Zealand which include survey methodology). Since a thorough professional knowledge is essential for both relevance and independence, most statistical offices wanting to maintain a strong methodology staff have no alternative to having a carefully designed career development program – whether methodology is organised in a centralised or decentralised manner.

For the courses to be relevant, it is desirable that a substantial portion of courses should be taught by staff members who are themselves active practitioners. This is easier arranged in centralised organisations where the senior methodologists can not only deploy staff to do teaching (typically on a part time basis), but can also arrange suitable replacements for them in their current project work.

The broader aspects of career development are also easier arranged in centralised organisations: they can more readily manage the periodic assignment of staff to different types of survey work, attendance at scientific conferences, the provision of research opportunities to those interested in and capable of doing part-time research work, and most importantly the service of apprenticeships under more experienced methodologists.

### *The case of Statistics Canada*

Training, not only in methodology, is emphasized by Statistics Canada (see Statistics Canada 1995). Overall, expenses on training amount to about 3% of its budget (or \$15 million) on formal training – plus a great deal more on various means of career development. But, in line with the centrality of training in methodology, the percentage of methodology budget spent on it is almost twice as much (bordering on 6 per cent in the 2008-09 fiscal year).

Training is provided in formal courses within Statistics Canada's Training Institute which currently (in 2009) offers some 20 courses in methodology, ranging in level from introductory courses to graduate level material. Most courses are taught by in-house staff, occasionally university personnel, mostly from local universities, are engaged if they are interested to teach and/or help develop our staff in other ways (*e.g.*, consultation) (in the latter modality we have been particularly fortunate in having had the contributions of Professor J.N.K. Rao of Carleton University over a period of some decades).



All recruits have to take a basic six weeks course which teaches (and provides practice in) survey design, survey operations, processing and analysis. This introductory training serves a multiplicity of purposes. Since the same basic six-week course in survey work is taken by *all* new professionals, it helps early on to inculcate in everyone a basic knowledge of all that is involved in survey work ; and, even more importantly, to drive home the critical importance of inter-disciplinary team work. It is also at this stage that new recruits from other disciplines are exposed for the first time to the requirements of methodology in survey design

Career development involves much more than training. The staff, particularly at the earlier stage of their career, is regularly given opportunities to work on different types of work: demographic, socio-economic, business surveys, use of administrative records, record linkage, *etc.* Significant numbers also attend scientific conferences. For example, during the last several years some 17 per cent of the methodology staff attended various Canadian and international professional conferences per annum. Staff is also encouraged to work on research projects and publish findings in peer reviewed journals, including Statistics Canada's *Survey Methodology*. Finally, for many years now Statistics Canada has organised an international methodology symposium to which leading research personnel from around the world are invited. These symposia are, of course, open to all Statistics Canada personnel and most methodologists choose to attend them.

## 8. Advisory Committee

### *General considerations*

A Methodology Advisory Committee can serve a most useful function (a) ensuring sound methodology practices, (b) integrating these practices into the daily work of statistical organisations, and (c) training staff. But the Committee can only be effective if (a) its advice is sought on significant issues of methodology and (b) there are mechanisms to ensure that the Committee's views are given due weight. I have observed Methodology Advisory Committees playing an equally useful role in a centralised office (Statistics Canada) and in a decentralised one (the Bureau of the Census in the 1960s).

### *The case of Statistics Canada*

Statistics Canada's Methodology Advisory Committee plays a key role. There are several factors that contribute to its usefulness and standing:

- The personal standing of the Committee's members is part of the reason.
- Every significant project of Statistics Canada is referred to the Committee for advice.
- The Committee's review is facilitated by the preparation of a paper for each item of the agenda which is introduced by a brief oral presentation by staff.
- Designated members of the Committee serve as formal discussants of each item on the agenda. The discussants present their views formally. Given that most of the papers are prepared by mid-career staff, these discussions make not only a substantive contribution to the projects that are discussed, but also to the training of the staff concerned – and that of the audience.
- Meetings of the Committee are attended not only by a large number of the relevant methodologists, but also by senior personnel of the subject matter division concerned, including often the Chief Statistician as well as one or two of his assistants.
- The Committee meets regularly: twice a year, for a day and a half on each occasion.
- The Committee regularly reviews the follow-up arising from its conclusions and formal recommendations; this helps ensure that their advice is taken seriously.

## 9. Research

### *General considerations*

I am taking it for granted that for this audience I do not need to spend time underscoring the intrinsic importance of research in a statistical agency. But let me stress the following points:

- Careful thought should be given to organising the research function in a manner that maximises both its relevance and the likelihood that its benefits will be successfully transmitted into daily practice. It is crucial to avoid the twin dangers of research being self-serving, or alternatively so completely task-oriented that it becomes pedestrian.
- Research needs to be adequately funded.
- In-house research needs to develop and to maintain close links with relevant extramural research.

### *The case of Statistics Canada*

One of the four methodology divisions is formally devoted to full time research. But the research is organised in a particular manner. Even though the research budget provides for the equivalent of 22 full time research staff, the research division itself has only six full time members. The remaining budget is assigned to finance the part-time research work of some other 70 methodologists. This arrangement serves a variety of purposes. First, it contributes to the relevance of research. Secondly, it



contributes to the adoption of the results of research. And thirdly, it helps morale for while not everyone wants to do research (or is able to do so), many want to try their hand at it. And the very act of conducting some research, by those capable of it, leads to more open mindsets and a better informed practice.

We are trying to ensure that the particular projects approved for research are in line with the broad research priorities of Statistics Canada, but at the same time leave some scope for self-initiated research. We do this by establishing broad priorities each year and inviting proposals in those areas from staff. The proposals are subject to formal adjudication: the best ones are selected and staff are assigned to work on them. Senior advice and guidance is provided by the Director of the Statistical Research and Innovation Division and its small permanent staff.

The following are additional measures that help the quality of research carried out:

- The possibility of publishing papers in *Survey Methodology*, Statistics Canada's own publication, serves as an incentive. While the peer review of the articles is rigorously managed by an international editorial board, the existence of a local yet prestigious outlet for methodology research represents a visible commitment by senior management.
- We regularly co-author papers with well known external research personnel (both Canadian and non-Canadian).
- We hold regular methodology interchanges with methodology staff in the US Bureaus of the Census and of Labour Statistics.
- We participate actively in Canadian, American and international statistical organisations.

## 10. Concluding comments

As indicated in the introduction, the bulk of the paper was devoted to the tools that should be considered by statistical offices in establishing and supporting the methodology function and the associated research, tools that in appropriate combination can enhance both the professional independence as well as the relevance of the function. I want to emphasise, however, that this is not a cook book. More important than all the tools is the environment: whether the statistical office welcomes

questioning and ensures that substantive questions are answered in substance; whether change is intrinsically frowned upon; whether it fosters collegiality; whether intelligent risk taking is encouraged or frowned upon; whether experiments are welcomed, assessed on their merits, and acted upon. These are the attributes that come from the top leadership of the statistical office and tools cannot substitute for them. Under the wrong leadership the best methodology staff (or, indeed, the best statistical office itself) will wither. But the contrary is not true: it is essential to have a careful understanding of the subtle balances advocated in this paper, as well as a careful deployment of the tools that give them effect. And even then, only a long term strategy can succeed.

I am completely certain that Joe would agree with my conclusion (see Waksberg 1998).

## References

- Brackstone, G.J. (1991). Shaping statistical services to satisfy user needs. *Statistical Journal of the United Nations Economic Commission for Europe*, 8, 3/4, 243-258.
- Brackstone, G. (1997). Organization of a survey methodology service. *Enquêtes et sondages : Méthodes, modèles, applications, nouvelles approches*, (Eds., G. Brossier and A.-M. Dussaix), Rennes, France, June 19-20, 3, 118-134.
- Fellegi, I.P. (1991). Maintaining public confidence in official statistics. *Journal of the Royal Statistical Society, Series A* (Statistics in Society), 154, Part 1, 1-6.
- Fellegi, I.P. (1992). Planning and Priority Setting - the Canadian Experience. In *Statistics in the Democratic Process at the End of the 20<sup>th</sup> Century*, (Eds., Hölder, Malaguerra and Vukovich); Anniversary publication for the 40<sup>th</sup> Plenary Session of the Conference of European Statisticians. Published by the Federal Statistical Office, Wiesbaden, Federal Republic of Germany.
- Fellegi, I.P. (1996). Characteristics of an effective statistical system. *International Statistical Review*, 64, 2, 165-199.
- Fellegi, I.P. (2004). Maintaining the credibility of official statistics. *Statistical Journal of the United Nations*, ECE 21, 191-198.
- Statistics Canada (1995). Training and Development at Statistics Canada. Statistics Canada Training Institute, March 1995.
- Waksberg, J. (1998). The Hansen Era: Statistical research and its implementation at the U.S. Census Bureau, 1940-1970. *Journal of Official Statistics*, 14, 2, 119-135.

# Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

Carl-Erik Särndal and Sixten Lundström<sup>1</sup>

## Abstract

This article develops computational tools, called indicators, for judging the effectiveness of the auxiliary information used to control nonresponse bias in survey estimates, obtained in this article by calibration. This work is motivated by the survey environment in a number of countries, notably in northern Europe, where many potential auxiliary variables are derived from reliable administrative registers for household and individuals. Many auxiliary vectors can be composed. There is a need to compare these vectors to assess their potential for reducing bias. The indicators in this article are designed to meet that need. They are used in surveys at Statistics Sweden. General survey conditions are considered: There is probability sampling from the finite population, by an arbitrary sampling design; nonresponse occurs. The probability of inclusion in the sample is known for each population unit; the probability of response is unknown, causing bias. The study variable (the  $y$ -variable) is observed for the set of respondents only. No matter what auxiliary vector is used in a calibration estimator (or in any other estimation method), a residual bias will always remain. The choice of a “best possible” auxiliary vector is guided by the indicators proposed in the article. Their background and computational features are described in the early sections of the article. Their theoretical background is explained. The concluding sections are devoted to empirical studies. One of these illustrates the selection of auxiliary variables in a survey at Statistics Sweden. A second empirical illustration is a simulation with a constructed finite population; a number of potential auxiliary vectors are ranked in order of preference with the aid of the indicators.

Key Words: Calibration weighting; Nonresponse adjustment; Nonresponse bias; Auxiliary variables; Bias indicator.

## 1. Introduction

Large nonresponse is typical of many surveys today. This creates a need for techniques for reducing as much as possible the nonresponse bias in the estimates. Powerful auxiliary information is needed. Administrative data files are a source of such information. The Scandinavian countries and some other European countries, notably the Netherlands, are in an advantageous position. Many potential auxiliary variables (called  $x$ -variables) can be taken from high quality administrative registers where auxiliary variable values are specified for the entire population. Variables measuring aspects of the data collection is another useful type of auxiliary data. Effective action can be taken to control nonresponse bias. Beyond sampling design, *design for estimation* becomes, in these countries, an important component of the total design. Statistics Sweden has devoted considerable resources to the development of techniques for selecting the best auxiliary variables.

Many articles discuss weighting in surveys with nonresponse and the selection of “best auxiliary variables”. Examples include Eltinge and Yansaneh (1997), Kalton and Flores-Cervantes (2003), and Thomsen, Kleven, Wang and Zhang (2006). Weighting in panel surveys with attrition receives special attention in, for example, Rizzo, Kalton and Brick (1996), who suggest that “the choice of auxiliary variables is an important one, and probably more important

than the choice of the weighting methodology”. The review by Kalton and Flores-Cervantes (2003) provides many references to earlier work. As in this paper, a calibration approach to nonresponse weighting is favoured in Deville (2002) and Kott (2006).

Some earlier methods are special cases of the outlook in this article, which is based on a systematic use of auxiliary information by calibration at two levels. Recently the search for efficient weighting has emphasized two directions: (i) to provide a more general setting than the popular but limited cell weighting techniques, and (ii) to quantify the search for auxiliary variables with the aid of computable indicators. Särndal and Lundström (2005, 2008) propose such indicators, while Schouten (2007) uses a different perspective to motivate an indicator. An article of related interest is Schouten, Cobben and Bethlehem (2009).

This content of this article has four parts: The general background for estimation with nonresponse is stated in Sections 2 to 4. Indicators for preference ranking of  $x$ -vectors are presented in Sections 5 and 6, and the computational aspects are discussed. The linear algebra derivations behind the indicators is presented in Sections 7 and 8. The two concluding Sections 9 and 10 present two empirical illustrations. The first (Section 9) uses real data from a large survey at Statistics Sweden. The second (Section 10) reports a simulation carried out on a constructed finite population.

1. Carl-Erik Särndal, Professor and Sixten Lundström, Senior Methodological Advisor, Statistics Sweden. E-mail: carl.sarndal@scb.se.



## 2. Calibration estimators for a survey with nonresponse

A probability sample  $s$  is drawn from the population  $U = \{1, 2, \dots, k, \dots, N\}$ . The sampling design gives unit  $k$  the known inclusion probability  $\pi_k = \Pr(k \in s) > 0$  and the known design weight  $d_k = 1/\pi_k$ . Nonresponse occurs. The response set  $r$  is a subset of  $s$ ; how it was generated is unknown. We assume  $r \subset s \subset U$ , and  $r$  non-empty. The (design weighted) response rate is

$$P = \frac{\sum_r d_k}{\sum_s d_k} \quad (2.1)$$

(if  $A$  is a set of units,  $A \subseteq U$ , a sum  $\sum_{k \in A}$  will be written as  $\sum_A$ ). Ordinarily a survey has many study variables. A typical one, whether continuous or categorical, is denoted  $y$ . Its value for unit  $k$  is  $y_k$ , recorded for  $k \in r$ , not available for  $k \in U - r$ . We seek to estimate the population  $y$ -total,  $Y = \sum_U y_k$ . Many parameters of interest in the finite population are functions of several totals, but we can focus on one such total.

The auxiliary information is of two kinds. To these correspond two vector types,  $\mathbf{x}_k^*$  and  $\mathbf{x}_k^\circ$ . *Population auxiliary information* is transmitted by  $\mathbf{x}_k^*$ , a vector value known for every  $k \in U$ . Thus  $\sum_U \mathbf{x}_k^*$  is a known population total. Alternatively, we allow that  $\sum_U \mathbf{x}_k^*$  is imported from an exterior source and that  $\mathbf{x}_k^*$  is a known (observed) vector value for every  $k \in s$ . *Sample auxiliary information* is transmitted by  $\mathbf{x}_k^\circ$ , a vector value known (observed) for every  $k \in s$ ; the total  $\sum_U \mathbf{x}_k^\circ$  is unknown but is estimated without bias by  $\sum_s d_k \mathbf{x}_k^\circ$ . The auxiliary vector value combining the two types is denoted  $\mathbf{x}_k$ . This vector and the associated information is

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}; \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}. \quad (2.2)$$

Tied to the  $k^{\text{th}}$  unit is the vector  $(y_k, \mathbf{x}_k, \pi_k)$ . Here,  $\pi_k$  is known for all  $k \in U$ ,  $y_k$  for all  $k \in r$ , the component  $\mathbf{x}_k^*$  of  $\mathbf{x}_k$  carries population information, the component  $\mathbf{x}_k^\circ$  of  $\mathbf{x}_k$  carries sample information.

Many  $\mathbf{x}$ -vectors can be formed with the aid of variables from administrative registers, survey process data or other sources. Among all the vectors at our disposal, we wish to identify the one most likely to reduce the nonresponse bias, if not to zero, so at least to a near-zero value.

We consider vectors having the property that there exists a constant non-null vector  $\boldsymbol{\mu}$  such that

$$\boldsymbol{\mu}' \mathbf{x}_k = 1 \text{ for all } k \in U \quad (2.3)$$

“Constant” means that  $\boldsymbol{\mu} \neq \mathbf{0}$  does not depend on  $k$ , nor on  $s$  or  $r$ . Condition (2.3) simplifies the mathematical derivations

and does not severely restrict  $\mathbf{x}_k$ . Most  $\mathbf{x}$ -vectors useful in practice are in fact covered. Examples include: (1)  $\mathbf{x}_k = (1, x_k)'$ , where  $x_k$  is the value for unit  $k$  of a continuous auxiliary variable  $x$ ; (2) the vector representing a categorical  $x$ -variable with  $J$  mutually exclusive and exhaustive classes,  $\mathbf{x}_k = \boldsymbol{\gamma}_k = (\gamma_{1k}, \dots, \gamma_{jk}, \dots, \gamma_{Jk})'$ , where  $\gamma_{jk} = 1$  if  $k$  belongs to group  $j$ , and  $\gamma_{jk} = 0$  if not,  $j = 1, 2, \dots, J$ ; (3) the vector  $\mathbf{x}_k$  used to codify two categorical variables, the dimension of  $\mathbf{x}_k$  being  $J_1 + J_2 - 1$ , where  $J_1$  and  $J_2$  are the respective number of classes, and the ‘minus-one’ is to avoid a singularity in the computation of weights calibrated to the two arrays of marginal counts; (4) the extension of (3) to more than two categorical variables. Vectors of the type (3) and (4) are especially important in statistics production in statistical agencies (the choice  $\mathbf{x}_k = x_k$ , not covered by (2.3), leads to the nonresponse ratio estimator, known to be a usually poor choice for controlling nonresponse bias, compared with  $\mathbf{x}_k = (1, x_k)'$ , so excluding the ratio estimator is no great loss).

The calibration estimator of  $Y = \sum_U y_k$ , computed on the data  $y_k$  for  $k \in r$ , is

$$\hat{Y}_{\text{CAL}} = \sum_r w_k y_k \quad (2.4)$$

with  $w_k = d_k \{1 + (\mathbf{X} - \sum_r d_k \mathbf{x}_k)(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k\}$ . The weights  $w_k$  are calibrated on both kinds of information:  $\sum_r w_k \mathbf{x}_k = \mathbf{X}$ , which implies  $\sum_r w_k \mathbf{x}_k^* = \sum_U \mathbf{x}_k^*$  and  $\sum_r w_k \mathbf{x}_k^\circ = \sum_s d_k \mathbf{x}_k^\circ$ . We assume throughout that the symmetric matrix  $\sum_r d_k \mathbf{x}_k \mathbf{x}_k'$  is nonsingular (for computational reasons, it is prudent to impose a stronger requirement: The matrix should not be ill-conditioned, or near-singular). In view of (2.3), we have  $\hat{Y}_{\text{CAL}} = \sum_r w_k y_k$  with weights  $w_k = d_k v_k$  where  $v_k = \mathbf{X}'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$ . The weights satisfy  $\sum_r d_k v_k \mathbf{x}_k = \mathbf{X}$ , where  $\mathbf{X}$  has one or both of the components in (2.2).

A closely related calibration estimator is based on the same two-tiered vector  $\mathbf{x}_k$  but with calibration only to the sample level:

$$\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k \quad (2.5)$$

where

$$m_k = \left( \sum_s d_k \mathbf{x}_k \right)' \left( \sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_k. \quad (2.6)$$

The calibration equation then reads  $\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$ , where  $\mathbf{x}_k$  has the two components as in (2.2). The auxiliary vector  $\mathbf{x}_k$  serves two purposes: To achieve a low variance and a low nonresponse bias. From the variance perspective alone,  $\hat{Y}_{\text{CAL}}$  is usually preferred to  $\tilde{Y}_{\text{CAL}}$  because the former profits from the input of a known population total  $\sum_U \mathbf{x}_k^*$ . But this paper studies the bias. From that perspective, we are virtually indifferent between  $\hat{Y}_{\text{CAL}}$  and



$\tilde{Y}_{\text{CAL}}$ , and we focus on the latter. Under liberal conditions, the difference between the bias of  $N^{-1}\hat{Y}_{\text{CAL}}$  and that of  $N^{-1}\tilde{Y}_{\text{CAL}}$  is of order  $n^{-1}$ , thereby of little practical consequence even for modest sample sizes  $n$ , as discussed for example in Särndal and Lundström (2005).

An alternative expression for (2.5) is

$$\tilde{Y}_{\text{CAL}} = \left( \sum_s d_k \mathbf{x}_k \right)' \mathbf{B}_x \quad (2.7)$$

where

$$\mathbf{B}_x = \mathbf{B}_{x|r;d} = \left( \sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_r d_k \mathbf{x}_k y_k \quad (2.8)$$

is the regression coefficient vector arising from the ( $d_k$ -weighted) least squares fit based on the data  $(y_k, \mathbf{x}_k)$  for  $k \in r$ .

A remark on the notation: When needed for emphasis, a symbol has two indices separated by a semicolon. The first shows the set of units over which the quantity is computed and the second indicates the weighting, as in  $\mathbf{B}_{x|r;d}$  given by (2.8), and in weighted means such as  $\bar{y}_{r;d} = \sum_r d_k y_k / \sum_r d_k$ . If the weighting is uniform, the second of the two indices is dropped as in  $\bar{y}_U = (1/N) \sum_U y_k$ .

### 3. Points of reference

The most primitive choice of vector is the constant one,  $\mathbf{x}_k = 1$  for all  $k$ . Although inefficient for reducing nonresponse bias, it serves as a benchmark. Then  $m_k = 1/P$  for all  $k$ , where  $P$  is the survey response rate (2.1), and  $\tilde{Y}_{\text{CAL}}$  is the expansion estimator:

$$\tilde{Y}_{\text{EXP}} = (1/P) \sum_r d_k y_k = \hat{N} \bar{y}_{r;d} \quad (3.1)$$

where  $\hat{N} = \sum_s d_k$  is design unbiased for the population size  $N$ . The bias of  $\tilde{Y}_{\text{EXP}}$  can be large.

At the opposite end of the bias spectrum are the unbiased, or nearly unbiased, estimators obtainable under full response, when  $r = s$ . They are hypothetical, not computable in the presence of nonresponse. Among these are the GREG estimator with weights calibrated to the known population total  $\sum_U \mathbf{x}_k^*$ ,

$$\hat{Y}_{\text{FUL}} = \sum_s d_k g_k y_k$$

where  $g_k = 1 + (\sum_U \mathbf{x}_k^* - \sum_s d_k \mathbf{x}_k^*)' (\sum_s d_k \mathbf{x}_k^* \mathbf{x}_k^{*'})^{-1} \mathbf{x}_k^*$ , and FUL refers to full response. The unbiased HT estimator (obtained when  $g_k = 1$  for all  $k$ ) is

$$\tilde{Y}_{\text{FUL}} = \sum_s d_k y_k = \hat{N} \bar{y}_{s;d}. \quad (3.2)$$

It disregards the information  $\sum_U \mathbf{x}_k^*$ , which may be important for variance reduction. But for the study of bias in this paper, we are indifferent between  $\hat{Y}_{\text{FUL}}$  and  $\tilde{Y}_{\text{FUL}}$ . The

difference in bias between the two is of little consequence, even for modest sample sizes. We can focus on  $\tilde{Y}_{\text{FUL}}$ .

### 4. The bias ratio

For a given outcome  $(s, r)$ , consider the estimates  $\tilde{Y}_{\text{CAL}}$ ,  $\tilde{Y}_{\text{EXP}}$  and  $\tilde{Y}_{\text{FUL}}$  given by (2.5), (3.1) and (3.2) as three points on a horizontal axis. Both  $\tilde{Y}_{\text{EXP}}$  (generated by the primitive  $\mathbf{x}_k = 1$ ) and  $\tilde{Y}_{\text{CAL}}$  (generated by a better  $\mathbf{x}$ -vector) are computable, but biased. As the  $\mathbf{x}$ -vector improves,  $\tilde{Y}_{\text{CAL}}$  will distance itself from  $\tilde{Y}_{\text{EXP}}$  and may come near the unbiased but unrealized ideal  $\tilde{Y}_{\text{FUL}}$ . We consider therefore three deviations:  $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$ ,  $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}$  and  $\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}$ , of which only the middle one is computable. The unknown “deviation total”,  $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$ , is decomposable as “deviation accounted for” (by the chosen  $\mathbf{x}$ -vector) plus “deviation remaining”:

$$\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}} = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}) + (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}). \quad (4.1)$$

If computable,  $\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}$  would be of particular interest, as an estimate of the bias remaining in  $\tilde{Y}_{\text{CAL}}$  (and in  $\hat{Y}_{\text{CAL}}$ ), whereas  $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$  would estimate the usually much larger bias of the benchmark,  $\tilde{Y}_{\text{EXP}}$ . The bias ratio for a given outcome  $(s, r)$  sets the estimated bias of  $\tilde{Y}_{\text{CAL}}$  in relation to that of  $\tilde{Y}_{\text{EXP}}$ :

$$\text{bias ratio} = \frac{\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}}{\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}}. \quad (4.2)$$

We scale the three deviations by the estimated population size  $\hat{N} = \sum_s d_k$  and use the notation  $\Delta_T = \Delta_A + \Delta_R$ , where  $T$  suggests “total”,  $A$  “accounted for” and  $R$  “remaining”. Noting that  $\sum_r d_k (y_k - \mathbf{x}_k' \mathbf{B}_x) = 0$ , we have

$$\Delta_T = \hat{N}^{-1} (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}) = \bar{y}_{r;d} - \bar{y}_{s;d};$$

$$\Delta_R = \hat{N}^{-1} (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) = \bar{\mathbf{x}}_{s;d}' \mathbf{B}_x - \bar{y}_{s;d}$$

$$\Delta_A = \hat{N}^{-1} (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}) = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_x$$

where  $\bar{\mathbf{x}}_{s;d} = \sum_s d_k \mathbf{x}_k / \sum_s d_k$ ,  $\bar{\mathbf{x}}_{r;d} = \sum_r d_k \mathbf{x}_k / \sum_r d_k$ , and  $\bar{y}_{s;d}$  and  $\bar{y}_{r;d}$  are the analogously defined means for the  $y$ -variable. Then (4.2) takes the form

$$\text{bias ratio} = \frac{\Delta_R}{\Delta_T} = 1 - \frac{\Delta_A}{\Delta_T} = 1 - \frac{(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_x}{\bar{y}_{r;d} - \bar{y}_{s;d}}. \quad (4.3)$$

We have bias ratio = 1 for the primitive vector  $\mathbf{x}_k = 1$ . Ideally, we want the auxiliary vector  $\mathbf{x}_k$  for  $\tilde{Y}_{\text{CAL}}$  to give bias ratio  $\approx 0$ . For a given outcome  $(s, r)$  and a given  $y$ -variable, we take steps in that direction by finding an  $\mathbf{x}$ -vector that makes the computable numerator  $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_x$  large (in absolute value). This is within our

reach. But whatever our final choice of  $\mathbf{x}$ -vector, the remaining bias of  $\tilde{Y}_{\text{CAL}}$  is unknown. Even with the best available  $\mathbf{x}$ -vector, considerable bias may remain. We have then attempted to do the best possible, under perhaps unfavourable circumstances.

To summarize, for a given outcome  $(s, r)$  and a given  $y$ -variable, the three deviations have the following features: (i)  $\Delta_T = \bar{y}_{r;d} - \bar{y}_{s;d}$  is an unknown constant value, depending on both unobserved and observed  $y$ -values; (ii)  $\Delta_A$  is computable; it depends on  $y_k$  for  $k \in r$  and on the values  $\mathbf{x}_k$  for  $k \in s$  of the chosen  $\mathbf{x}$ -vector; (iii)  $\Delta_R$  cannot be computed; it depends on unobserved values  $y_k$ , and on  $\mathbf{x}_k$  for  $k \in s$ .

To follow the progression of the estimates when the  $\mathbf{x}$ -vector improves, consider a given outcome  $(s, r)$ . The deviation  $\Delta_T$  can have either sign. Suppose  $\Delta_T > 0$ , indicating a positive bias in  $\tilde{Y}_{\text{EXP}}$ , as when large units respond with greater propensity than small ones. When the  $\mathbf{x}$ -vector in  $\tilde{Y}_{\text{CAL}}$  becomes progressively more powerful by the inclusion of more and more  $x$ -variables,  $\Delta_A$  tends to increase away from zero and will, ideally, come near  $\Delta_T$ , indicating a desired closeness of  $\tilde{Y}_{\text{CAL}}$  to the unbiased  $\tilde{Y}_{\text{FUL}}$ . As long as the  $\mathbf{x}$ -vector remains relatively weak,  $\Delta_A < \Delta_T$  is likely to hold. When the  $\mathbf{x}$ -vector becomes increasingly powerful,  $\Delta_A$  moves closer to the fixed  $\Delta_T$ , a sign of bias nearing zero. It could even “move beyond”, so that an “over-adjustment”,  $\Delta_A > \Delta_T$ , has occurred. This not a detrimental feature; although  $\Delta_R = \Delta_T - \Delta_A$  is then negative, it is ordinarily small (the analyst can only work with  $\Delta_A$ ; it is unknown to him/her whether  $\Delta_A$  and  $\Delta_T$  are close, or whether the over-adjustment  $\Delta_A > \Delta_T$  has occurred). These points are illustrated by the simulation in Section 10. If  $\Delta_T < 0$ , these tendencies are reversed.

The form of (4.3) may suggest an argument which can however be misleading: Suppose that a vector  $\mathbf{x}_k$  has been suggested, containing variables thought to be effective, along with an assumption that  $y_k = \boldsymbol{\beta}'\mathbf{x}_k + \varepsilon_k$ , where  $\varepsilon_k$  is a small residual. Then  $\bar{y}_{r;d} - \bar{y}_{s;d} \approx (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_x \approx (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\boldsymbol{\beta}$ , and consequently bias ratio  $\approx 0$ , sending a message, often false, that the postulated vector  $\mathbf{x}_k$  is efficient. One weakness of the argument stems from the well-known fact that nonresponse (unless completely random) will cause  $\mathbf{B}_x$  to be biased for a regression vector that describes the  $y$ -to- $\mathbf{x}$  relationship in the population. Further comments on this issue are given in Section 8.

Finally, there is the practical consideration that a typical survey has many  $y$ -variables. To every  $y$ -variable corresponds a calibration estimator, and a bias ratio given by (4.3). The ideal  $\mathbf{x}$ -vector is one that would be capable of controlling bias in all those estimators. This is usually not possible without compromise, as we discuss later.

## 5. Expressing the deviation accounted for

The responding unit  $k$  receives the weight  $d_k m_k$  in the estimator  $\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k$ . The nonresponse adjustment factor  $m_k = (\sum_s d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$  expands the design weight  $d_k$ . We can view  $m_k$  as the value of a derived variable, defined for a particular outcome  $(r, s)$  and choice of  $\mathbf{x}_k$ , independent of all  $y$ -variables of interest, and computable for  $k \in s$  (but used in  $\tilde{Y}_{\text{CAL}}$  only for  $k \in r$ ). Using (2.3), we have

$$\begin{aligned} \sum_r d_k m_k \mathbf{x}_k &= \sum_s d_k \mathbf{x}_k; \quad \sum_r d_k m_k = \sum_s d_k; \\ \sum_r d_k m_k^2 &= \sum_s d_k m_k. \end{aligned} \quad (5.1)$$

Two weighted means are needed:

$$\bar{m}_{r;d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k} = \frac{1}{P}; \quad \bar{m}_{s;d} = \frac{\sum_s d_k m_k}{\sum_s d_k} \quad (5.2)$$

where  $P$  is the response rate (2.1). Thus the average adjustment factor in  $\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k$  is  $1/P$ , regardless of the choice of  $\mathbf{x}$ -vector. Whether a chosen  $\mathbf{x}$ -vector is efficient or not for reducing bias will depend on higher moments of the  $m_k$ . The weighted variance of the  $m_k$  is

$$S_m^2 = S_{m|r;d}^2 = \sum_r d_k (m_k - \bar{m}_{r;d})^2 / \sum_r d_k. \quad (5.3)$$

The simpler notation  $S_m^2$  will be used. A development of (5.3) and a use of (5.1) and (5.2) gives

$$S_m^2 = \bar{m}_{r;d} (\bar{m}_{s;d} - \bar{m}_{r;d}). \quad (5.4)$$

The coefficient of variation of the  $m_k$  is

$$\text{cv}_m = \frac{S_m}{\bar{m}_{r;d}} = \sqrt{\frac{\bar{m}_{s;d}}{\bar{m}_{r;d}} - 1}. \quad (5.5)$$

The weighted variance of the study variable  $y$  is given by

$$S_y^2 = S_{y|r;d}^2 = \sum_r d_k (y_k - \bar{y}_{r;d})^2 / \sum_r d_k \quad (5.6)$$

(when the response probabilities are not all equal,  $S_y^2 = S_{y|r;d}^2$  is not unbiased for the population variance  $S_{y|U}^2$ , but this is not an issue for the derivations that follow). We need the covariance

$$\text{Cov}(y, m) = \text{Cov}(y, m)_{r;d} =$$

$$\frac{1}{\sum_r d_k} \sum_r d_k (m_k - \bar{m}_{r;d})(y_k - \bar{y}_{r;d}) \quad (5.7)$$

and the correlation coefficient,  $R_{y,m} = \text{Cov}(y, m) / (S_y S_m)$ , satisfying  $-1 \leq R_{y,m} \leq 1$ .

The deviation  $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_x$  is a crucial component in the bias ratio (4.3). We seek an  $\mathbf{x}$ -vector that



makes  $\Delta_A$  large. The factors that determine  $\Delta_A$  are seen in (5.8) to (5.10). Computational tools (indicators) to assist the search for effective  $x$ -variables are given in (5.11) and (5.12). Their derivation by linear algebra is deferred to Section 7, which may be bypassed by readers more interested in the practical use of these tools in the search for  $x$ -variables, as illustrated in the empirical Sections 9 and 10. We can factorize  $\Delta_A/S_y$  as

$$\Delta_A/S_y = -R_{y,m} \times cv_m. \quad (5.8)$$

Two simple multiplicative factors determine  $\Delta_A/S_y$ : The coefficient of variation  $cv_m$ , which is free of  $y_k$  and computed on the known  $\mathbf{x}_k$  alone, and the (positive or negative) correlation coefficient  $R_{y,m}$ . Another factorization in terms of simple concepts is

$$\Delta_A/S_y = F \times R_{y,x} \times cv_m \quad (5.9)$$

where  $R_{y,x} = \sqrt{R_{y,x}^2}$  is the coefficient of multiple correlation between  $y$  and  $\mathbf{x}$ ,  $R_{y,x}^2$  is the proportion of the  $y$ -variance  $S_y^2$  explained by the predictor  $\mathbf{x}$ , and  $F = -R_{y,m}/R_{y,x}$  (formula (7.8) states the precise expression for  $R_{y,x}^2$ ). As Section 7 also shows,  $|R_{y,m}| \leq R_{y,x}$  for any  $\mathbf{x}$ -vector and  $y$ -variable; consequently  $-1 \leq F \leq 1$ .

In (5.8) and (5.9),  $cv_m$  and  $R_{y,x}$  are non-negative terms, while  $R_{y,m}$  and  $F$  can have either sign (or possibly be zero). Hence

$$|\Delta_A|/S_y = |R_{y,m}| \times cv_m = |F| \times R_{y,x} \times cv_m. \quad (5.10)$$

All of  $S_y$ ,  $cv_m$ ,  $R_{y,x}$ ,  $R_{y,m}$  and  $F$  are easily computed in the survey. Both  $cv_m$  and  $R_{y,x}$  increase (or possibly stay unchanged) when further  $x$ -variables are added to the  $\mathbf{x}$ -vector;  $R_{y,m}$  does not have this property.

To illustrate with the aid of fairly typical numbers, if  $F = 0.5$ ;  $R_{y,x} = 0.6$  and  $cv_m = 0.4$ , then  $\Delta_A/S_y = 0.12$ , implying that  $\tilde{Y}_{\text{CAL}}/N = \tilde{Y}_{\text{EXP}}/N - 0.12 \times S_y$ . That is, the estimated  $y$ -mean  $\tilde{Y}_{\text{CAL}}/\hat{N}$  has become adjusted by 0.12 standard deviations down from the primitive estimate  $\tilde{Y}_{\text{EXP}}/\hat{N}$ . The adjustment can be large compared to the standard deviation of the estimated  $y$ -mean, especially when the survey sample size is in the thousands. It remains unknown whether or not that adjustment has cured most of the biasing effect of nonresponse.

It follows from (5.8) that  $0 \leq |\Delta_A|/S_y \leq cv_m$  whatever the  $y$ -variable. A sharper inequality is  $|\Delta_A|/S_y \leq R_{y,x} \times cv_m$ , but it depends on the  $y$ -variable. Further, if the correlation ratio  $F$  stays roughly constant when the  $\mathbf{x}$ -vector changes, so that  $F \approx F_0$ , then  $|\Delta_A|/S_y \approx |F_0| \times R_{y,x} \times cv_m$ .

Although computable for any  $\mathbf{x}$ -vector and any outcome  $(s, r)$ ,  $\Delta_A$  does not reveal the value of the bias ratio. But  $\Delta_A$  suggests computational tools, called indicators, for comparing alternative  $\mathbf{x}$ -vectors. By (5.8), let

$$H_0 = \Delta_A/S_y = -R_{y,m} \times cv_m. \quad (5.11)$$

As borne out by theory in Section 8 and by the empirical work in Section 10, over a long run of outcomes  $(s, r)$ , the average of  $H_0$  tracks the average deviation  $\tilde{Y}_{\text{CAL}} - Y$  (which measures the bias of  $\tilde{Y}_{\text{CAL}}$ ) in a nearly perfect linear manner when the  $\mathbf{x}$ -vector changes. This holds independently of the response distribution that generates  $r$  from  $s$ . Since  $H_0$  can have either sign, it is practical to work with its absolute value denoted  $H_1$ ; in addition we consider two other indicators,  $H_2$  and  $H_3$ , inspired by (5.9) to (5.10):

$$\begin{aligned} H_1 &= |\Delta_A|/S_y = |R_{y,m}| \times cv_m; \\ H_2 &= R_{y,x} \times cv_m; \quad H_3 = cv_m. \end{aligned} \quad (5.12)$$

Our main alternatives are  $H_1$  and  $H_3$ . Of these,  $H_1$  is motivated by its direct link to  $\Delta_A$ , which we want to make large, for a given  $y$ -variable. A strong reason to consider  $H_3$  is its independence of all  $y$ -variables in the survey. The indicator  $H_2$  is an *ad hoc* alternative; although  $H_2$  contains a familiar concept, the multiple correlation coefficient  $R_{y,x}$ , it is less appropriate than  $H_1$  because the correlation coefficient ratio  $F = -R_{y,m}/R_{y,x}$  may vary considerably from one  $\mathbf{x}$ -vector to another. Both  $H_2$  and  $H_3$  increase when further  $x$ -variables are added to the  $\mathbf{x}$ -vector, something which does not hold in general for  $H_1$ . The use of these indicators is illustrated in the empirical Sections 9 and 10.

## 6. Preference ranking of auxiliary vectors

The methods in this paper are intended for use primarily with the large samples that characterize government surveys. The sample size is ordinarily much larger than the dimension of the  $\mathbf{x}$ -vector. The variance of estimates is ordinarily small, compared to the squared bias. However, for categorical auxiliary variables, no group size should be allowed to be “too small”. It is recommended that all group sizes be at least 30, if not at least 50, in order to avoid instability. The crossing of categorical variables (to allow interactions) implies a certain risk of small groups. It is preferable to calibrate on marginal counts, rather than on frequencies for small crossed cells.

In a number of countries, the many available administrative registers provide a rich source of auxiliary information, particularly for surveys on individuals and households. These registers contain many potential  $x$ -variables from which to choose. Many different  $\mathbf{x}$ -vectors can be composed. The indicators in (5.12) provide computational tools for obtaining a preference ordering, or a ranking, of potential  $\mathbf{x}$ -vectors, with the objective to reduce



as much as possible the bias remaining in the calibration estimator.

*Scenario 1:* Focus on a specific  $y$ -variable. The bias remaining in the calibration estimator depends on the  $y$ -variable; some are more bias prone than others. We identify one specific  $y$ -variable deemed to be highly important in the survey, and we seek to identify an  $\mathbf{x}$ -vector that reduces the bias for this variable as much as possible (if more than one  $y$ -variable needs to be taken into account, a compromise must be struck, which suggests Scenario 2 below). For this purpose, we use the  $y$ -variable dependent indicator  $H_1 = |\Delta_A|/S_y = |R_{y,m}| \times cv_m$  and choose the  $\mathbf{x}$ -vector so as to make  $H_1$  large. An *ad hoc* alternative is to use the indicator  $H_2 = R_{y,x} \times cv_m$ , and strive to make it as large as possible.

*Scenario 2:* The objective is to identify a general purpose  $\mathbf{x}$ -vector, efficient for all or most  $y$ -variables in the survey. This suggests  $H_3 = cv_m$  as a compromise indicator, and to choose the  $\mathbf{x}$ -vector that maximizes  $H_3$ . To that same effect, Särndal and Lundström (2005, 2008) used the indicator  $S_m^2 = H_3^2 / P^2$ . They showed that the derived variable  $m_k$  in (2.6) can be seen as a predictor of the inverse of the unknown response probability and that choosing the  $\mathbf{x}$ -vector to make  $S_m^2$  large signals a bias reduction in the calibration estimator, irrespective of the  $y$ -variable.

For each scenario we can distinguish two procedures:

*All vectors procedure:* A list of candidate  $\mathbf{x}$ -vectors is prepared, based on appropriate judgment. We compute the chosen indicator for *every* candidate  $\mathbf{x}$ -vector, and settle for the vector that gives the highest indicator value. The resulting  $\mathbf{x}$ -vector may not be the same for  $H_1$  (which targets a specific  $y$ -variable) as for  $H_3$  (which seeks a compromise for all  $y$ -variables in the survey).

*Stepwise procedure:* There is a pool of available  $x$ -variables. We build the  $\mathbf{x}$ -vector by a stepwise forward (or stepwise backward) selection from among the available  $x$ -variables, one variable at a time, using the successive changes (if considered large enough) in the value of the chosen indicator to signal the inclusion (or exclusion) of a given  $x$ -variable at a given step. The indicators  $H_1$ ,  $H_2$  and  $H_3$  do not in general give the same selection of variables. Consider two  $\mathbf{x}$ -vectors,  $\mathbf{x}_{1k}$  and  $\mathbf{x}_{2k}$ , such that  $\mathbf{x}_{2k}$  is made up of  $\mathbf{x}_{1k}$  and an additional vector  $\mathbf{x}_{+k}$ :  $\mathbf{x}_{2k} = (\mathbf{x}_{1k}, \mathbf{x}_{+k})'$ . The transition from  $\mathbf{x}_{1k}$  to  $\mathbf{x}_{2k}$  will increase the value of  $H_2$  and  $H_3$ . In each step of a forward selection procedure we select the variable bringing the largest increase in  $H_2$  or  $H_3$ . But the transition does not guarantee an increased value for the most appropriate indicator,  $H_1$ . However,  $H_1$  may be used in stepwise selection in the manner described in Section 9.

## 7. Derivations

For given  $y$ -variable and outcome  $(s, r)$ , we seek an  $\mathbf{x}$ -vector to make the computable numerator  $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_x$  in the bias ratio (4.3) large, in absolute value. In this section we prove the factorizations  $\Delta_A/S_y = -R_{y,m} \times cv_m = F \times R_{y,x} \times cv_m$  in (5.8) and (5.9). We note first that  $cv_m^2$  is a quadratic form in the vector that contrasts the  $\mathbf{x}$ -mean in the response set  $r$  with the  $\mathbf{x}$ -mean in the sample  $s$ . Let

$$\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d}; \quad \Sigma = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' / \sum_r d_k. \quad (7.1)$$

Then, with  $P$  given by (2.1),

$$cv_m^2 = P^2 \times S_m^2 = \mathbf{D}' \Sigma^{-1} \mathbf{D}. \quad (7.2)$$

This expression follows from (5.3) and a consequence of (2.3), namely,

$$\bar{\mathbf{x}}_{r;d}' \Sigma^{-1} \bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}_{r;d}' \Sigma^{-1} \bar{\mathbf{x}}_{s;d} = 1. \quad (7.3)$$

The vector of covariances with the study variable  $y$  is

$$\mathbf{C} = \left( \sum_r d_k (\mathbf{x}_k - \bar{\mathbf{x}}_{r;d}) (y_k - \bar{y}_{r;d}) \right) / \left( \sum_r d_k \right). \quad (7.4)$$

We can then write  $\Delta_A$  as a bilinear form:

$$\Delta_A = \mathbf{D}' \mathbf{B}_x = \mathbf{D}' \Sigma^{-1} \mathbf{C} \quad (7.5)$$

using that  $\mathbf{D}' \Sigma^{-1} \bar{\mathbf{x}}_{r;d} = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \Sigma^{-1} \bar{\mathbf{x}}_{r;d} = 0$  by (7.3).

A useful perspective on  $\Delta_A$  is gained from the geometric interpretation of  $\mathbf{C}$  and  $\mathbf{D}$  in (7.5) as vectors in the space whose dimension is that of  $\mathbf{x}_k$ . We have

$$\Delta_A = \Lambda (\mathbf{D}' \Sigma^{-1} \mathbf{D})^{1/2} (\mathbf{C}' \Sigma^{-1} \mathbf{C})^{1/2} \quad (7.6)$$

where

$$\Lambda = \frac{\mathbf{D}' \Sigma^{-1} \mathbf{C}}{(\mathbf{D}' \Sigma^{-1} \mathbf{D})^{1/2} (\mathbf{C}' \Sigma^{-1} \mathbf{C})^{1/2}}. \quad (7.7)$$

For a specific  $y$ -variable and a specific  $\mathbf{x}$ -vector, the scalar quantities  $(\mathbf{D}' \Sigma^{-1} \mathbf{D})^{1/2}$  and  $(\mathbf{C}' \Sigma^{-1} \mathbf{C})^{1/2}$  represent the respective vector lengths of  $\mathbf{D}$  and  $\mathbf{C}$  (following an orthogonal transformation based on the eigenvectors and eigenvalues of  $\Sigma^{-1}$ ). The scalar quantity  $\Lambda$  represents the cosine of the angle between  $\mathbf{D}$  (which is independent of  $y$ ) and  $\mathbf{C}$  (which depends on  $y$ ); hence  $-1 \leq \Lambda \leq 1$ .

When the auxiliary vector  $\mathbf{x}_k$  is allowed to expand by adding further available  $x$ -variables, both vector lengths  $(\mathbf{D}' \Sigma^{-1} \mathbf{D})^{1/2}$  and  $(\mathbf{C}' \Sigma^{-1} \mathbf{C})^{1/2}$  increase. The change in the angle  $\Lambda$  may be in either direction; if  $|\Lambda|$  stays roughly constant, (7.6) shows that  $|\Delta_A|$  will increase.

A second useful perspective on  $\Delta_A$  follows by decomposing the total variability of the study variable  $y$ ,  $\sum_r d_k (y_k - \bar{y}_{r;d})^2 = (\sum_r d_k) S_y^2$ . Two regression fits need

to be examined, the one of  $y$  on the auxiliary vector  $\mathbf{x}$ , and the one of  $y$  on the derived variable  $m$  defined by (2.6). To each fit corresponds a decomposition of  $S_y^2$  into explained  $y$ -variation and residual  $y$ -variation. The two explained portions have important links to the bias ratio (4.3). Result 7.1 summarizes the two decompositions.

**Result 7.1.** For a given survey outcome  $(s, r)$ , let  $\mathbf{D}$ ,  $\Sigma$  and  $\mathbf{C}$  be given by (7.1) and (7.4). Then the proportion of the  $y$ -variance  $S_y^2$  explained by the regression of  $y$  on  $\mathbf{x}$  is

$$R_{y,\mathbf{x}}^2 = (\mathbf{C}'\Sigma^{-1}\mathbf{C})/S_y^2. \quad (7.8)$$

The coefficient of correlation between  $y$  and the univariate predictor  $m$  is

$$R_{y,m} = -(\mathbf{D}'\Sigma^{-1}\mathbf{C})/[(\mathbf{D}'\Sigma^{-1}\mathbf{D})^{1/2} \times S_y]. \quad (7.9)$$

Consequently, the proportion of  $S_y^2$  explained by  $m$  is

$$R_{y,m}^2 = (\mathbf{D}'\Sigma^{-1}\mathbf{C})^2 / [(\mathbf{D}'\Sigma^{-1}\mathbf{D}) \times S_y^2]. \quad (7.10)$$

The proportions  $R_{y,\mathbf{x}}^2$  and  $R_{y,m}^2$  satisfy  $R_{y,m}^2 \leq R_{y,\mathbf{x}}^2 \leq 1$ .

*Proof.* The proof of (7.8) uses the weighted least squares regression of  $y$  on  $\mathbf{x}$  fitted over  $r$ . The residuals are  $y_k - \hat{y}(\mathbf{x})_k$ , where  $\hat{y}(\mathbf{x})_k = \mathbf{x}'_k \mathbf{B}_\mathbf{x}$  with  $\mathbf{B}_\mathbf{x}$  given by (2.8). The decomposition is

$$\begin{aligned} \sum_r d_k (y_k - \bar{y}_{r;d})^2 &= \sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 \\ &\quad + \sum_r d_k (y_k - \hat{y}(\mathbf{x})_k)^2. \end{aligned}$$

The mixed term is zero. A development of the term “variation explained” gives  $\sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 = (\sum_r d_k) \mathbf{C}'\Sigma^{-1}\mathbf{C}$ . Thus the proportion of variance explained is  $R_{y,\mathbf{x}}^2 = \sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 / [(\sum_r d_k) S_y^2] = \mathbf{C}'\Sigma^{-1}\mathbf{C} / S_y^2$ , as claimed in (7.8). To show (7.9) we note that the covariance (5.7) can be written with the aid of (7.5) as

$$\text{Cov}(y, m) = -\Delta_A / P = -\mathbf{D}'\Sigma^{-1}\mathbf{C} / P.$$

It then follows from (7.2) that  $R_{y,m} = \text{Cov}(y, m) / (S_y S_m)$  has the expression (7.9). The residuals from the regression (with intercept) of  $y$  on the univariate explanatory variable  $m$  are  $\hat{y}(m)_k = \bar{y}_{r;d} + B_m(m_k - \bar{m}_{r;d})$  with  $B_m = \text{Cov}(y, m) / S_m^2 = -P(\mathbf{D}'\Sigma^{-1}\mathbf{C}) / (\mathbf{D}'\Sigma^{-1}\mathbf{D})$ . The proportion of variance explained is  $\sum_r d_k (\hat{y}(m)_k - \bar{y}_{r;d})^2 / [(\sum_r d_k) S_y^2]$ , which upon development gives the expression for  $R_{y,m}^2$  in (7.10). Finally,  $R_{y,m}^2 \leq R_{y,\mathbf{x}}^2$  follows from the Cauchy-Schwarz inequality for a bilinear form:  $(\mathbf{D}'\Sigma^{-1}\mathbf{C})^2 \leq (\mathbf{D}'\Sigma^{-1}\mathbf{D})(\mathbf{C}'\Sigma^{-1}\mathbf{C})$ .

The inequality  $R_{y,m}^2 \leq R_{y,\mathbf{x}}^2 \leq 1$  can also be deduced by the fact that, among all predictions  $\hat{y}_k = \mathbf{x}'_k \beta$  that are linear in the  $\mathbf{x}$ -vector, those that maximize the variance explained are  $\hat{y}(\mathbf{x})_k = \mathbf{x}'_k \mathbf{B}_\mathbf{x}$ , so the predictions  $\hat{y}(m)_k$ , which are

linear in  $\mathbf{x}_k$  via  $m_k$ , cannot yield a greater variance explained than that maximum.

Now from (7.9), (7.2) and (7.5),  $-R_{y,m} \text{cv}_m = \mathbf{D}'\Sigma^{-1}\mathbf{C} / S_y = \Delta_A / S_y$ , as claimed by formula (5.8). Moreover, (7.7), (7.8) and (7.9) imply  $-R_{y,m} / R_{y,\mathbf{x}} = \Lambda$ , so the correlation coefficient ratio  $F$  in (5.9) equals the angle  $\Lambda$  defined by (7.7).

## 8. Comments: Goodness of fit, properties of the bias and a related selection procedure

Three issues are examined in this section: (i) The relationship between bias and goodness of fit, (ii) the linear relation between the expected value of  $\Delta_A = \hat{N}^{-1}(\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}})$  and the bias of  $\tilde{Y}_{\text{CAL}}$  or  $\hat{Y}_{\text{CAL}}$ , and (iii) the alternative method for selection of auxiliary variables proposed by Schouten (2007).

For the issue (i), recall that the total deviation in Section 4 is  $\Delta_T = \Delta_A + \Delta_R$ , where  $\Delta_A$  is computable but  $\Delta_T$  and  $\Delta_R$  are not. If computable,  $\hat{N} \Delta_R = \tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}$  would be an estimate of the bias of  $\tilde{Y}_{\text{CAL}}$  (and of that of  $\hat{Y}_{\text{CAL}}$ ). A small  $\Delta_R$  is desirable. The question arises: Is this achieved when  $y_k = \beta' \mathbf{x}_k + \varepsilon_k$  (with a given vector  $\mathbf{x}_k$ ) fits the data well? We need to distinguish two aspects: (a) The computable fit to the data  $(y_k, \mathbf{x}_k)$  observed for  $k \in r$ ; and (b) The hypothetical fit to the data  $(y_k, \mathbf{x}_k)$  for  $k \in s$ , some observed, some not.

A good fit for the respondents,  $k \in r$ , does not guarantee a small  $\Delta_R$ : The weighted LSQ fit using the observed data  $(y_k, \mathbf{x}_k)$  for  $k \in r$  gives the residuals  $e_{k|r;d} = y_k - \mathbf{x}'_k \mathbf{B}_{\mathbf{x}|r;d}$ , computable for  $k \in r$ , with the property  $\sum_r d_k e_{k|r;d} = 0$  (here, the detailed notation  $\mathbf{B}_{\mathbf{x}|r;d}$  specified in (2.8) is preferable to the simplified notation  $\mathbf{B}_\mathbf{x}$ ). For  $k \in s - r$ ,  $e_{k|r;d}$  is not computable; it has an unknown non-zero mean  $\bar{e}_{s-r;d} = \sum_{s-r} d_k e_{k|r;d} / \sum_{s-r} d_k$ . We have

$$\Delta_R = (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) / \hat{N} = -(1 - P) \bar{e}_{s-r;d} \neq 0. \quad (8.1)$$

Regardless of whether the fit is good (small residuals  $e_{k|r;d}$ ;  $R_{y,\mathbf{x}}^2$  near one) or poor (large residuals  $e_{k|r;d}$ ;  $R_{y,\mathbf{x}}^2$  near zero), the deviation  $\Delta_R$  given by (8.1) may be large, and  $\tilde{Y}_{\text{CAL}}$  far from unbiased. Even with a perfect fit for the respondents ( $e_{k|r;d} = 0$  for all  $k \in r$ , and  $R_{y,\mathbf{x}}^2 = 1$ ), there is no guarantee that the bias is small.

A similar inadequacy affects imputation based on the respondent data. If the regression imputations  $\hat{y}_k = \mathbf{x}'_k \mathbf{B}_{\mathbf{x}|r;d}$  are used to fill in for the values  $y_k$  missing for  $k \in s - r$ , the imputed estimator is

$$\hat{Y}_{\text{imp}} = \sum_r d_k y_k + \sum_{s-r} d_k \hat{y}_k.$$

Then  $\hat{Y}_{\text{imp}} = \tilde{Y}_{\text{CAL}}$ , so  $\hat{Y}_{\text{imp}}$  has the same exposure to bias as  $\tilde{Y}_{\text{CAL}}$ , as is easily understood: When the nonresponse



causes a skewed selection of  $y$ -values, the imputed values computed on that skewed selection will misrepresent the unknown  $y$ -values that characterize the sample  $s$  or the population  $U$ .

Consider now the aspect (b) of the fit, that is, the hypothetical weighted LSQ regression fit to the data  $(y_k, \mathbf{x}_k)$  for  $k \in s$ . The regression coefficient vector would be  $\mathbf{B}_{\mathbf{x}|s;d} = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_s d_k \mathbf{x}_k y_k$ , and the residuals  $e_{k|s;d} = y_k - \mathbf{x}_k' \mathbf{B}_{\mathbf{x}|s;d}$  for  $k \in s$  satisfy  $\sum_s d_k e_{k|s;d} = 0$ . Using that  $\sum_r d_k m_k \mathbf{x}_k / \hat{N} = \bar{\mathbf{x}}_{s;d}$  and  $\sum_r d_k m_k y_k / \hat{N} = \bar{y}_{s;d}$ , we have

$$\Delta_R = \hat{N}^{-1}(\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) = (1/\hat{N}) \sum_r d_k m_k e_{k|s;d}. \quad (8.2)$$

Suppose the model is “true for the sample  $s$ ”, with a perfect fit, so that  $e_{k|s;d} = 0$  for all  $k \in s$ . Then, by (8.2) we do have  $\Delta_R = 0$ , so the nonresponse adjusted estimator  $\tilde{Y}_{\text{CAL}}$  agrees with the unbiased estimator  $\tilde{Y}_{\text{FUL}}$ . A belief that the bias is small hinges on an unverifiable assumption.

Turning to the issue (ii), we now explain the essentially linear relation between the bias of  $\tilde{Y}_{\text{CAL}}$  and the expected value of the indicator  $H_0 = \Delta_A/S_y = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}})/\hat{N}S_y$ . For a given outcome  $(s, r)$ , a fixed  $y$ -variable and a fixed  $\mathbf{x}$ -vector we have

$$(\tilde{Y}_{\text{CAL}} - Y)/\hat{N}S_y = (\tilde{Y}_{\text{EXP}} - Y)/\hat{N}S_y - H_0.$$

Let  $E_{pq}$  denote the expectation operator with respect to all outcomes  $(s, r)$ , that is,  $E_{pq}(\cdot) = E_p(E_q(\cdot|s))$ , where  $p(s)$  and  $q(r|s)$  are, respectively, the known sampling design and the unknown response distribution. We denote  $\text{bias}(\tilde{Y}_{\text{CAL}}) = E_{pq}(\tilde{Y}_{\text{CAL}}) - Y$ ,  $\text{bias}(\tilde{Y}_{\text{EXP}}) = E_{pq}(\tilde{Y}_{\text{EXP}}) - Y$  and  $C = E_{pq}(\hat{N}S_y)$ . Using the usual large sample replacement of the expected value of a ratio by the ratio of the expected values, we have  $E_{pq}[(\tilde{Y}_{\text{CAL}} - Y)/\hat{N}S_y] \approx [E_{pq}(\tilde{Y}_{\text{CAL}}) - Y]/E_{pq}(\hat{N}S_y)$  and analogously for  $\tilde{Y}_{\text{EXP}}$ , so

$$\text{bias}(\tilde{Y}_{\text{CAL}}) \approx \text{bias}(\tilde{Y}_{\text{EXP}}) - C \times E(H_0). \quad (8.3)$$

Here  $\text{bias}(\tilde{Y}_{\text{EXP}})$  and  $C$  do not depend on the choice of  $\mathbf{x}$ -vector, whereas  $\text{bias}(\tilde{Y}_{\text{CAL}})$  and  $E(H_0)$  do. Therefore, as the  $\mathbf{x}$ -vector changes,  $\text{bias}(\tilde{Y}_{\text{CAL}})$  and  $E(H_0)$  are essentially linearly related. No particular forms of  $p(s)$  and  $q(r|s)$  need to be specified for (8.3) to hold. As a consequence, when two auxiliary vectors,  $\mathbf{x}_{1k}$  and  $\mathbf{x}_{2k}$ , are compared, the difference in bias is, to close approximation, proportional to the change in the expected value of  $H_0$ :

$$\text{bias}(\tilde{Y}_{\text{CAL}}(\mathbf{x}_{1k})) - \text{bias}(\tilde{Y}_{\text{CAL}}(\mathbf{x}_{2k})) \approx -C(E_1 - E_2) \quad (8.4)$$

where  $E_i = E_{pq}(H_0(\mathbf{x}_{ik}))$  for  $i = 1, 2$ . The properties (8.3) and (8.4) are validated by the Monte Carlo study in Section 10.

Note that formula (8.3) does not guarantee that  $\tilde{Y}_{\text{CAL}}$  based on a certain vector  $\mathbf{x}_k$  will have zero or near-zero bias. It does not state that a comparatively large value of  $|\Delta_A|$  guarantees a small bias in  $\tilde{Y}_{\text{CAL}}$ . What (8.3) says is that  $\text{bias}(\tilde{Y}_{\text{CAL}})$  is linearly related to the expectation of the indicator  $H_0 = \Delta_A/S_y$ . Therefore, to assess available  $\mathbf{x}$ -vectors in terms of the indicator  $H_0$  (or the indicator  $H_1 = |\Delta_A|/S_y$ ) is consistent with the objective of bias reduction.

Turning to the issue (iii), we comment on the alternative method for selection of auxiliary variables proposed by Schouten (2007). His indicator for the step-by-step selection of variables differs from our indicators; it will usually not select exactly the same set of variables. In a list of say 30 available categorical  $\mathbf{x}$ -variables, the first ten to enter will not be the same set of ten as with our indicators  $H_0$  to  $H_3$ . The order in which variables are selected will not necessarily be the same either. For comparison, we compared, in some of our empirical work, with the variable selection realized by Schouten's method. In some cases we noted a considerable congruence between the two sets of “first ten” picked in the two procedures.

The differences between the two approaches are best appreciated by a comparison of their background and derivation. Our indicators  $H_0$  and  $H_1$  originate in the notion of separation (or distance), for a given outcome  $(s, r)$ , between the adjusted estimator  $\tilde{Y}_{\text{CAL}}$  and the primitive one,  $\tilde{Y}_{\text{EXP}}$ , and in the idea that this separation will ordinarily increase when the  $\mathbf{x}$ -vector becomes more powerful. The probability sampling design is taken into consideration; no assumptions are made on the response distribution.

Schouten uses a superpopulation argument; sampling weights do not appear to enter into consideration. An expression for the model-expected bias of an estimator of the population mean is found to be proportional to the correlation (at the level of the population) between the  $y$ -variable and the 0-1 indicator for response. It is shown that this correlation (and consequently the bias) can be bounded inside an interval. In particular, the generalized regression estimator is considered and it is shown that its maximum absolute bias equals the width of the bias interval. This width depends on the true unknown regression vector  $\beta$  for the regression (at the population level) between  $y$  and  $\mathbf{x}$ . This unknown  $\beta$  is replaced by an estimate based on the respondents, thus subject to some bias because of the nonresponse. Schouten emphasizes that a missing-at-random assumption is not needed for his method, which is in that respect similar to our method.



## 9. Auxiliary variable choice for the Swedish pilot survey on gaming and problem gambling

We identified a real survey data set to illustrate the use of the indicators  $H_1$ ,  $H_2$  and  $H_3$  in building the  $\mathbf{x}$ -vector. In 2008, The Swedish National Institute of Public Health (*Svenska Folkhälsoinstitutet*) conducted a pilot survey to study the extent of gambling participation and the characteristics of persons with gambling problems. Sampling and weight calibration was carried out by Statistics Sweden. We illustrate the use of the indicators in this survey, for which a stratified simple random sample  $s$  of  $n = 2,000$  persons was drawn from the Swedish Register of Total Population (RTP). The strata were defined by the cross classification of region of residence by age group. Each of the six regions was defined as a cluster of postal code areas deemed similar in regard to variables such as education level, purchasing power, type of housing, foreign background. The four age groups were defined by the brackets 16-24; 25-34; 35-64 and 65-84.

The overall unweighted response rate was 50.8%. The nonresponse, more or less pronounced in the different domains of interest, interferes with the accuracy objective. An extensive pool of potential auxiliary variables was available for this survey, including variables in the RTP, in the Education Register and a subset of those in another extensive Statistics Sweden data base, LISA. For this illustration, we prepared a data file consisting of 13 selected categorical variables. Twelve of these were designated as  $x$ -variables, and one, the dichotomous variable *Employed*, played the role of the study variable. The values of all variables are available for all units  $k \in s$ . Response ( $k \in r$ ) or not ( $k \in s - r$ ) to the survey is also indicated in the data file.

Variables that are continuous by nature were used as grouped; all 12  $x$ -variables are thus categorical and of the  $\mathbf{x}_k^\circ$  type, as defined in Section 2 (because most of the variables are available for the full population, they are potentially of the type  $\mathbf{x}_k^*$ , but since the effect on bias is of little consequence, we used them as  $\mathbf{x}_k^\circ$ -variables). The study variable value,  $y_k = 1$  if  $k$  is *employed* and  $y_k = 0$  otherwise, is known for  $k \in s$ , so the unbiased estimate  $\tilde{Y}_{\text{FUL}}$  defined by (3.2) can be computed and used as a reference. We also computed  $\tilde{Y}_{\text{EXP}}$  defined by (3.1), as well as  $\tilde{Y}_{\text{CAL}}$  defined by (2.5) for different  $\mathbf{x}$ -vectors built by stepwise selection from the pool of 12  $x$ -variables with the aid of the indicators  $H_1$ ,  $H_2$  and  $H_3$  defined by (5.12).

We carried out forward selection as follows: The auxiliary vector in Step 0 is the trivial  $\mathbf{x}_k = 1$ , and the estimator is  $\tilde{Y}_{\text{EXP}}$ . In Step 1, the indicator value is computed for every one of 12 presumptive auxiliary variables; the variable producing the largest value of the indicator is

selected. In Step 2, the indicator value is computed for all 11 vectors of dimension two that contain the variable selected in Step 1 and one of the remaining variables. The variable that gives the largest value for the indicator is selected in Step 2, and so on, in the following steps. A new variable always joins already entered variables in the “side-by-side” (or “+”) manner. Interactions are thereby relinquished. The order of selection is different for each indicator.

The values of  $H_2$  and  $H_3$  that identify the next variable for inclusion are by mathematical necessity increasing in every step. This does not hold for  $H_1$ . In a certain step  $j$ , we used the rule to include the  $x$ -variable with the largest of computed  $H_1$ -values. That value can be smaller than the  $H_1$ -value that identified the variable entering in the preceding step,  $j - 1$ . The series of  $H_1$ -values for inclusion will increase up to a certain step, then begin to decline, as Table 9.1 illustrates.

The unbiased estimate is  $\tilde{Y}_{\text{FUL}} = 4,265$ ; the primitive estimate is  $\tilde{Y}_{\text{EXP}} = 4,719$  (both in thousands). This suggests a large positive bias in  $\tilde{Y}_{\text{EXP}}$ , whose relative deviation (in %) from  $\tilde{Y}_{\text{FUL}}$  is  $\text{RDF} = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}) / \tilde{Y}_{\text{FUL}} \times 10^2 = 10.7$ . Adding categorical  $x$ -variables one by one into the  $\mathbf{x}$ -vector will successively change this deviation, although when a few variables have been admitted, the change is not always in the direction of a smaller value. In each step we computed the indicator,  $\tilde{Y}_{\text{CAL}}$  and  $\text{RDF} = (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) / \tilde{Y}_{\text{FUL}} \times 10^2$ .

Table 9.1 shows the stepwise selection with the indicator  $H_1$  (the number of categories is given in parenthesis for each selected variable). First to enter is the variable *Income class*; this brings a large reduction in RDF from 10.7 to 4.5. The next five selections take place with increased  $H_1$ -values, and the value of RDF is reduced, but by successively smaller amounts. Step six, where *Marital status* is selected, brings about a turning point, indicated by the double line in Table 9.1: The value of  $H_1$  then starts to decline, and  $\tilde{Y}_{\text{CAL}}$  and RDF start to increase. At step 6, RDF is at its lowest value, 0.5, then starts to rise, illustrating that inclusion of all available  $x$ -variables may not be best. The turning point of  $H_1$  and the point at which RDF is closest to zero happen to agree in this example. This is not generally the case. Moreover, in a real survey setting, RDF is unknown, as is the step at which RDF is closest to zero.

Table 9.2 shows the stepwise selection with indicator  $H_3$ . Its value increases at every step, but at a rate that levels off, and successive changes in  $\tilde{Y}_{\text{CAL}}$  become negligible. This suggests to stop after six steps, at which point  $\text{RDF} = 2.8$ . In none of the 12 steps does RDF come as close to zero as the value  $\text{RDF} = 0.5$  obtained with  $H_1$  after six steps. In this respect  $H_1$  is better than  $H_3$ , in this example. With all 12  $x$ -variables selected, RDF attains in both tables the final value 2.6.

Table 9.1

Stepwise forward selection, indicator  $H_1$ , dichotomous study variable *Employed*. Successive values of  $H_1 \times 10^3$ , of  $\tilde{Y}_{\text{CAL}}$  in thousands, and of  $\text{RDF} = (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}})/\tilde{Y}_{\text{FUL}} \times 10^2$ . For comparison,  $\tilde{Y}_{\text{EXP}} \times 10^{-3} = 4,719$ ;  $\tilde{Y}_{\text{FUL}} \times 10^{-3} = 4,265$

Auxiliary variable entered	$H_1 \times 10^3$	$\tilde{Y}_{\text{CAL}} \times 10^{-3}$	RDF
Income class (3)	76	4,458	4.5
Education level (3)	107	4,350	2.0
Presence of children (2)	114	4,326	1.4
Urban centre dwelling (2)	118	4,310	1.1
Sex (2)	123	4,296	0.7
Marital status (2)	125	4,286	0.5
Days unemployed (3)	121	4,301	0.9
Months with sickness benefits (3)	120	4,305	1.0
Level of debt (3)	115	4,322	1.3
Cluster of postal codes (6)	109	4,343	1.8
Country of birth (2)	103	4,363	2.3
Age class (4)	99	4,377	2.6

Table 9.2

Stepwise forward selection, indicator  $H_3$ , dichotomous study variable *Employed*. Successive values of  $H_3 \times 10^3$ , of  $\tilde{Y}_{\text{CAL}}$  in thousands, of  $\text{RDF} = (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}})/\tilde{Y}_{\text{FUL}} \times 10^2$ . For comparison,  $\tilde{Y}_{\text{EXP}} \times 10^{-3} = 4,719$ ;  $\tilde{Y}_{\text{FUL}} \times 10^{-3} = 4,265$

Auxiliary variable entered	$H_3 \times 10^3$	$\tilde{Y}_{\text{CAL}} \times 10^3$	RDF
Education level (3)	186	4,520	6.0
Cluster of postcode areas (6)	250	4,505	5.6
Country of birth (2)	281	4,498	5.5
Income Class (3)	298	4,369	2.4
Age class (4)	354	4,399	3.1
Sex (2)	364	4,384	2.8
Urban centre dwelling (2)	374	4,378	2.6
Level of debt (3)	381	4,364	2.3
Months with sickness benefits (3)	384	4,380	2.7
Presence of children (2)	387	4,379	2.7
Marital status (2)	388	4,379	2.7
Days unemployed (3)	388	4,377	2.6

The set of the first six variables to enter with  $H_3$  has three in common with the corresponding set of six with  $H_1$ . There is no contradiction in the quite different selection patterns, because  $H_1$  is geared to the specific  $y$ -variable *Employed*, while  $H_3$  is a compromise indicator, independent of any  $y$ -variable. To save space, the step-by-step results for indicator  $H_2$  are not shown. Its selection pattern resembles more that of  $H_3$  than that of  $H_1$ . Out of the first six variables to enter with  $H_2$ , four are among the first six with  $H_3$ . As a general comment, we believe that in many practical situations the use of more than six variables is unnecessary, and the selection of the first few becomes crucially important.

## 10. Empirical validation by simulation for a constructed population

The theory presented in earlier sections makes no assumptions on the response distribution. It is unknown. The sampling design is arbitrary; its known inclusion

probabilities are taken into account. For the experiment in this section, we specify several different response distributions with a specified positive value for the response probability  $\theta_k$  for every  $k \in U$ . That is, with specified probability  $\theta_k$ , the value  $y_k$  gets recorded in the experiment; with probability  $1 - \theta_k$ , it goes missing. We find that the indicators  $H_0$  (or  $H_1 = |H_0|$ ) defined in (5.11) ranks the different  $\mathbf{x}$ -vectors in the correct order of preference for all participating response distributions, consistent with the theoretical results (8.3) and (8.4). We confirm that, over a long run of outcomes  $(s, r)$ , the average of  $H_0 = \Delta_A/S_y = -R_{y,m} \times \text{cv}_m$  tracks the bias of the calibration estimator, measured by the average of  $\tilde{Y}_{\text{CAL}} - Y$ , in an essentially perfectly linear manner, when the  $\mathbf{x}$ -vector moves through 16 different formulations. We also examine the indicators  $H_2$  and  $H_3$  defined in (5.12), and find in this experiment that they also have strong relationship to the bias of  $\tilde{Y}_{\text{CAL}}$ .

We experimented with several created populations; the conclusions were similar. We report here results for one constructed population of size  $N = 6,000$ , with created values  $(y_k, \mathbf{x}_k, \theta_k)$  for  $k = 1, 2, \dots, N = 6,000$ , for 16 alternative categorical formulations of  $\mathbf{x}_k$ , and four different ways to assign the  $\theta_k$ .

The 16 alternative categorical auxiliary  $\mathbf{x}$ -vectors were obtained by grouping the generated values  $x_{1k}$  and  $x_{2k}$  of two continuous auxiliary variables,  $x_1$  and  $x_2$ . The values  $(y_k, x_{1k}, x_{2k})$  for  $k = 1, 2, \dots, 6,000$  were created in three steps as follows. Step 1 (the variable  $x_1$ ): The 6,000 values  $x_{1k}$  were obtained as independent outcomes of the gamma distributed random variable  $\Gamma(a, b)$  with parameter values  $a = 2, b = 5$ . The mean and variance of the 6,000 realized values  $x_{1k}$  was 10.0 and 49.9, respectively. Step 2 (the variable  $x_2$ ): For unit  $k$ , with value  $x_{1k}$  fixed by Step 1, a value  $x_{2k}$  is realized as an outcome of the gamma random variable with parameters such that the conditional expectation and variance of  $x_{2k}$  are  $\alpha + \beta x_{1k} + K h(x_{1k})$  and  $\sigma^2 x_{1k}$ , respectively, where  $h(x_{1k}) = x_{1k}(x_{1k} - \mu_{x_1}) (x_{1k} - 3\mu_{x_1})$  with  $\mu_{x_1} = 10$ . We used the values  $\alpha = 1, \beta = 1, K = 0.001$  and  $\sigma^2 = 25$ . The polynomial term  $K h(x_{1k})$  gives a mild non-linear shape to the plot of  $(x_{2k}, x_{1k})$ , to avoid an exactly linear relationship. The mean and variance of the 6,000 realized values  $x_{2k}$  were 11.0 and 210.0, respectively. The correlation coefficient between  $x_1$  and  $x_2$ , computed on the 6,000 couples  $(x_{1k}, x_{2k})$ , was 0.48. Step 3 (the study variable  $y$ ): For unit  $k$ , with values  $x_{1k}$  and  $x_{2k}$  fixed by Steps 1 and 2, a value  $y_k$  is realized as an outcome of the gamma random variable with parameters such that the conditional expectation and variance of  $y_k$  are  $c_0 + c_1 x_{1k} + c_2 x_{2k}$  and  $\sigma_0^2 (c_1 x_{1k} + c_2 x_{2k})$ , respectively. We used  $c_0 = 1, c_1 = 0.7, c_2 = 0.3$  and  $\sigma_0^2 = 2$ . The mean and the variance of the 6,000 realized



values  $y_k$  were 11.4 and 86.5, respectively. The correlation coefficient between  $y$  and  $x_1$ , computed on the 6,000 couples  $(y_k, x_{1k})$ , was 0.76; that between  $y$  and  $x_2$ , computed on the 6,000 couples  $(y_k, x_{2k})$ , was 0.73.

Each of the two  $x$ -variables was then transformed into four alternative group modes, denoted 8G, 4G, 2G and 1G, yielding  $4 \times 4 = 16$  different auxiliary vectors  $\mathbf{x}_k$ . The 6,000 values  $x_{1k}$  of variable  $x_1$  were size ordered; eight equal-sized groups were formed. Group 1 consists of the units with the 750 largest values  $x_{1k}$ , group 2 consists of the next 750 units in the size ordering, and so on, ending with group 8. In this mode 8G of  $x_1$ , unit  $k$  is assigned the vector value  $\gamma_{(x_1;8)k}$ , of dimension eight with seven entries "0" and a single entry "1" to code the group membership of  $k$ . Next, successive group mergers are carried out, so that two adjoining groups always define a new group, every time doubling the group size. Thus for mode 4G, the merger of groups 1 and 2 puts the units with the 1,500 largest  $x_{1k}$ -values into a first new group; groups 3 and 4 merge to form the second new group of 1,500, and so on; the vector value associated with unit  $k$  is  $\gamma_{(x_1;4)k}$ . In mode 2G, unit  $k$  has the vector value  $\gamma_{(x_1;2)k} = (1, 0)'$  for the 3,000 largest  $x_1$ -value units and  $\gamma_{(x_1;2)k} = (0, 1)'$  for the rest. In the ultimate mode, 1G, all 6,000 units are put together, all  $x_1$ -information is relinquished, and  $\gamma_{(x_1;1)k} = 1$  for all  $k$ . The 6,000 values  $x_{2k}$  were transformed by the same procedure into the group modes 8G, 4G, 2G and 1G. Corresponding group membership of unit  $k$  is coded by the vectors  $\gamma_{(x_2;8)k}$ ,  $\gamma_{(x_2;4)k}$ ,  $\gamma_{(x_2;2)k}$  and  $\gamma_{(x_2;1)k} = 1$ . The  $4 \times 4 = 16$  different auxiliary vectors  $\mathbf{x}_k$  take into account both kinds of group information; the two  $\gamma$ -vectors are placed side by side (as opposed to crossed), the result being a calibration on two margins, as indicated by the "+" sign. Thus for the case denoted 8G + 8G, unit  $k$  has the auxiliary vector value  $\mathbf{x}_k = (\gamma'_{(x_1;8)k}, \gamma'_{(x_2;8)k})'_{(-1)}$ , where  $(-1)$  indicates that one category is excluded in either  $\gamma_{(x_1;8)k}$  or  $\gamma_{(x_2;8)k}$  to avoid a singular matrix in the computation, giving  $\mathbf{x}_k$  the dimension  $8 + 8 - 1 = 15$ . The case 8G + 8G has the highest information content. At the other extreme, the case 1G + 1G disregards all the  $x$ -information and  $\mathbf{x}_k = 1$  for all  $k$ . There are 14 intermediate cases of information content. For example, 4G + 2G has  $\mathbf{x}_k = (\gamma'_{(x_1;4)k}, \gamma'_{(x_2;2)k})'_{(-1)}$  of dimension  $4 + 2 - 1 = 5$ ; 4G + 1G has  $\mathbf{x}_k = (\gamma'_{(x_1;4)k}, 1)'_{(-1)} = \gamma_{(x_1;4)k}$  of dimension 4 (there is non-negligible interaction between  $x_1$  and  $x_2$  in this experiment, but we restrict the experiment to  $\mathbf{x}$ -vectors without interactions, causing no risk of small group counts).

We discuss here the results for four response distributions. Their response probabilities  $\theta_k$ ,  $k = 1, 2, \dots, N = 6,000$ , were specified as follows:

$$\text{IncExp}(10 + x_1 + x_2), \quad \text{with } \theta_k = 1 - e^{-c(10 + x_{1k} + x_{2k})} \\ \text{where } c = 0.04599$$

$$\begin{aligned} &\text{IncExp}(10 + y), && \text{with } \theta_k = 1 - e^{-c(10 + y_k)} \\ & && \text{where } c = 0.06217 \\ &\text{DecExp}(x_1 + x_2), && \text{with } \theta_k = e^{-c(x_{1k} + x_{2k})} \\ & && \text{where } c = 0.01937 \\ &\text{DecExp}(y), && \text{with } \theta_k = e^{-cy_k} \\ & && \text{where } c = 0.03534. \end{aligned}$$

The constant  $c$  was adjusted in all four cases to give a mean response probability of  $\bar{\theta}_U = \sum_U \theta_k / N = 0.70$ . In the first two, the value 10 (rather than 0) was used to avoid a high incidence of small response probabilities  $\theta_k$ . These four options represent contrasting features for the response probabilities: increasing as opposed to decreasing, dependent on  $x$ -values only as opposed to dependent on  $y$ -values only. In the second and fourth option, the response is directly  $y$ -variable dependent, and could hence be called "purely non-ignorable".

We generated  $J = 5,000$  outcomes  $(s, r)$ , where  $s$  of size  $n = 1,000$  is drawn from  $N = 6,000$  by simple random sampling and, for every given  $s$ , the response set  $r$  is realized by each of the four response distributions. That is, for  $k \in s$ , a Bernoulli trial was carried out with the specified probability  $\theta_k$  of inclusion in the response set  $r$ . The Bernoulli trials are independent.

For each response distribution, for each of the 16  $\mathbf{x}$ -vectors, and for every outcome  $(s, r)$ , we computed the relative deviation  $\text{RD} = (\hat{Y}_{\text{CAL}} - Y) / Y$ , where  $\hat{Y}_{\text{CAL}}$  is given by (2.4) and  $Y = \sum_U y_k$  is the targeted  $y$ -total, known in this experimental setting (alternatively, we used  $\tilde{Y}_{\text{CAL}}$  given by (2.5) but, as expected, the difference in bias compared with  $\hat{Y}_{\text{CAL}}$  is negligible). We also computed the indicators  $H_i$ ,  $i = 0, 1, 2, 3$ , given by (5.11) and (5.12). Summary measures were computed as

$$\text{relbias} = \text{Av}(\text{RD}) = \frac{1}{J} \sum_{j=1}^J \text{RD}_j;$$

$$\text{Av}(H_i) = \frac{1}{J} \sum_{j=1}^J H_{ij} \quad \text{for } i = 0, 1, 2, 3$$

where  $j$  indicates the value computed for the  $j^{\text{th}}$  outcome,  $j = 1, 2, \dots, 5,000 = J$ . For each response distribution, we thus obtain the value *relbias* (which is the Monte Carlo measure of the relative bias  $(E_{pq}(\hat{Y}_{\text{CAL}}) - Y) / Y$ ) and 16 values of  $\text{Av}(H_i)$  (which is the Monte Carlo measure of  $E_{pq}(H_i)$ ),  $i = 0, 1, 2, 3$ , where  $p$  stands for simple random sampling, and  $q$  stands for one of the four response distributions.

Table 10.1 shows, for  $\text{IncExp}(10 + x_1 + x_2)$ , *relbias* in % and  $\text{Av}(H_i) \times 10^3$  for the 16  $\mathbf{x}$ -vectors. For the cell 1G + 1G, with vector  $\mathbf{x}_k = 1$ , all four *Av*-quantities are zero, and *relbias* is at its highest level, 13.2%. At the opposite extreme, the cell 8G + 8G represents the highest level of



information; it gives the highest value for  $Av(H_1)$ , and  $relbias$  is at its lowest value, 0.2%; virtually all bias is removed (except for a possible sign difference,  $Av(H_0)$  and  $Av(H_1)$  were equal for all cells).

The result (8.4), holding for any response distribution and any sampling design, states that the indicator  $H_0$  will rank the  $4 \times 4 = 16$  auxiliary vectors correctly for any response distribution (with response probabilities not all constant, as noted below). Table 10.1 illustrates (8.4) in terms of  $H_1 = |H_0|$ : The change, from any one cell to any other, in the value of  $Av(H_1)$  (the Monte-Carlo estimate of the expected value of  $(H_1)$  is accompanied by a proportional change in the value of  $relbias$ . The same proportionality was noted for the other three response distributions. We could have chosen other response distributions to illustrate the same property.

**Table 10.1**  
Relbias in % and, within parenthesis, the value of  $Av(H_1) \times 10^3$  for 16 auxiliary vectors  $x_k$ . Response distribution IncExp(10 +  $x_1 + x_2$ )

Groups based on $x_{1k}$	Groups based on $x_{2k}$							
	8G		4G		2G		1G	
8G	0.2	(101)	0.5	(99)	1.3	(93)	3.4	(76)
4G	0.5	(98)	0.9	(96)	1.8	(89)	4.1	(70)
2G	1.5	(91)	1.9	(88)	3.2	(78)	6.5	(52)
1G	4.1	(70)	5.0	(64)	7.3	(46)	13.2	(0)

The response distribution with a constant response probability  $\theta_k$  for all  $k$  is a special case. The calibration estimator  $\tilde{Y}_{CAL}$  based on any vector  $x_k$  then has zero bias (very nearly), and this includes the primitive estimator  $\tilde{Y}_{EXP}$  with  $x_k = 1$ . Result 8.3 continues to be valid, stating in that case that  $E_{pq}(H_0) \approx \text{bias}(\tilde{Y}_{CAL}) \approx \text{bias}(\tilde{Y}_{EXP}) \approx 0$ . In the context of the simulation in this section, if  $\theta_k = 0.70$  for all  $k$  is taken to be an additional response distribution, Table 10.1 will in all 16 cells show nearly zero values of both  $relbias$  in % and  $Av(H_1) \times 10^3$ , from the weakest cell (1G + 1G) all the way to the cell of the most powerful  $x$ -vector (8G + 8G). There is no bias to be removed by an improvement of the  $x$ -vector. If in practice the indicator ( $H_1$ ) does not react to an enlargement of the  $x$ -vector, there is no incentive to seek beyond the simplest vector formulation. It could signify one of three possibilities: The  $y$ -variable in question is not subject to nonresponse bias, or that the response probability is almost constant, or that none of the available  $x$ -vectors is capable of reducing an existing bias.

To save space we do not show the corresponding tables for  $Av(H_2)$  and  $Av(H_3)$ . By mathematical necessity, both quantities increase in the nested transitions. Not shown either are the counterparts of Table 10.1 for the other three response distributions. The patterns are similar.

Table 10.2 for IncExp(10 +  $x_1 + x_2$ ) and Table 10.3 for IncExp(10 +  $y$ ) show how  $Av(H_1)$ ,  $Av(H_2)$  and  $Av(H_3)$  rank the 16  $x$ -vectors, represented by their value of  $relbias$ . To measure the success of ranking, we computed the Spearman rank correlation coefficient, denoted  $rancor$ , between  $relbias$  and the value of the indicator, based on the 16 values of each. For  $Av(H_1)$ , the bottom line of the two tables shows  $|rancor| = 1$ , for perfect ranking. For these data,  $|rancor|$  is near one also for  $Av(H_2)$  and  $Av(H_3)$  (more generally, the ranking obtained with  $H_2$  and  $H_3$  may be good, but is data dependent).

**Table 10.2**  
Value, in ascending order, of  $relbias$  in %, and corresponding value and rank of  $Av(H_1) \times 10^3$ ,  $Av(H_2) \times 10^3$  and  $Av(H_3) \times 10^3$ , for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlations,  $rancor$ . Response distribution IncExp(10 +  $x_1 + x_2$ )

relbias	$Av(H_1) \times 10^3$		$Av(H_2) \times 10^3$		$Av(H_3) \times 10^3$	
0.2	101	(1)	127	(1)	232	(1)
0.5	99	(2)	119	(2)	225	(2)
0.5	98	(3)	118	(3)	224	(3)
0.8	96	(4)	109	(4)	217	(4)
1.3	93	(5)	109	(5)	216	(5)
1.5	91	(6)	105	(6)	213	(6)
1.8	89	(7)	98	(7)	207	(7)
1.9	88	(8)	94	(8)	205	(8)
3.2	78	(9)	80	(11)	192	(9)
3.4	76	(10)	90	(9)	188	(11)
4.1	70	(11)	84	(10)	190	(10)
4.1	70	(12)	77	(12)	175	(13)
5.0	64	(13)	70	(13)	179	(12)
6.4	52	(14)	52	(14)	146	(15)
7.3	46	(15)	46	(15)	156	(14)
13.2	0	(16)	0	(16)	0	(16)
<i>Rancor</i>	-1.00		-0.99		-0.99	

There is one notable contrast between the results on  $relbias$  for the two response distributions in Tables 10.2 and 10.3. The best among the auxiliary vectors leave considerably more bias for the non-ignorable IncExp(10 +  $y$ ) than for IncExp(10 +  $x_1 + x_2$ ). This is not unexpected, and it is important to note that considerable bias reduction is obtained for the non-ignorable case as well.

In the simulation, the over-adjustment mentioned in Section 4,  $\Delta_A > \Delta_T > 0$  (when  $(\tilde{Y}_{EXP})$  has positive bias) or  $\Delta_A < \Delta_T < 0$  (when  $\tilde{Y}_{EXP}$  has negative bias), happens for some outcomes  $(s, r)$ . The frequency varies with the strength of the auxiliary vector and is different for different response distributions. The cell for which this over-adjustment is most likely to occur is 8G + 8G, the most powerful of the 16 auxiliary vectors. For IncExp(10 +  $x_1 + x_2$ ), the bias is almost completely removed for cell 8G + 8G;  $relbias$  is only 0.2%. Hence  $\tilde{Y}_{CAL}$  is close to the unbiased  $\tilde{Y}_{FUL}$ ,  $\Delta_A$  is near  $\Delta_T$ , and  $\Delta_A > \Delta_T$  happened for 45.6% of all outcome  $(s, r)$ . By contrast, for the non-ignorable case IncExp(10 +  $y$ ), the incidence of  $\Delta_A > \Delta_T$

was only 0.1% for the cell  $8G + 8G$ . Although that cell brings considerable bias reduction (compared to the primitive  $1G + 1G$ ), there is bias remaining, and as a consequence,  $\Delta_A > \Delta_T$  almost never happens.

We do not show the corresponding tables for  $\text{DecExp}(x_1 + x_2)$  and  $\text{DecExp}(y)$ . The lowest value of *rancor* was 0.94, recorded for  $\text{Av}(H_3)$  in the case of  $\text{DecExp}(x_1 + x_2)$ .

A question not addressed in Tables 10.2 and 10.3 is: How often, over a long series of outcomes  $(s, r)$ , does a given indicator  $H(\mathbf{x}_k)$  succeed in pointing correctly to the preferred  $\mathbf{x}$ -vector? To answer this, let  $\mathbf{x}_{1k}$  and  $\mathbf{x}_{2k}$  be two vectors selected for comparison. If the absolute value of the bias of  $\hat{Y}_{\text{CAL}}(\mathbf{x}_{2k})$  is smaller than that of  $\hat{Y}_{\text{CAL}}(\mathbf{x}_{1k})$ , we would like to see that  $H(\mathbf{x}_{2k}) \geq H(\mathbf{x}_{1k})$  holds for a vast majority of all outcomes  $(s, r)$ , because then the indicator  $H(\cdot)$  delivers with high probability the correct decision to prefer  $\mathbf{x}_{2k}$ . Because  $H(\mathbf{x}_k)$  has sampling variability, its success rate (the rate of correct indication) depends on the sample size, and we expect it to increase with sample size.

Table 10.3

Value, in ascending order, of *relbias* in %, and corresponding value and rank of  $\text{Av}(H_1) \times 10^3$ ,  $\text{Av}(H_2) \times 10^3$  and  $\text{Av}(H_3) \times 10^3$ , for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlations, *rancor*. Response distribution  $\text{IncExp}(10 + y)$

<i>relbias</i>	$\text{Av}(H_1) \times 10^3$		$\text{Av}(H_2) \times 10^3$		$\text{Av}(H_3) \times 10^3$	
3.6	74	(1)	91	(1)	165	(1)
3.9	71	(2)	84	(2)	158	(2)
4.0	71	(3)	83	(3)	156	(3)
4.3	68	(4)	76	(5)	149	(5)
4.4	68	(5)	78	(4)	153	(4)
4.9	64	(6)	68	(8)	142	(3)
4.9	63	(7)	72	(6)	146	(6)
5.3	60	(8)	69	(7)	143	(7)
5.4	60	(9)	64	(9)	137	(9)
6.0	55	(10)	59	(10)	132	(10)
6.2	53	(11)	54	(11)	128	(11)
7.2	46	(12)	54	(12)	122	(12)
7.9	41	(13)	41	(14)	111	(13)
7.9	40	(14)	43	(13)	109	(14)
9.6	27	(15)	27	(15)	90	(15)
13.1	0	(16)	0	(16)	0	(16)
<i>Rancor</i>	-1.00		-0.99		-0.99	

We threw some light on this question by extending the Monte Carlo experiment: 5,000 outcomes  $(s, r)$  were realized, first with sample size  $n = 1,000$ , then with sample size  $n = 2,000$  (the response set  $r$  is realized according to one of the four response distributions, declaring unit  $k$  “responding” as a result of a Bernoulli trial with the specified probability  $\theta_k$ ). We computed the success rate as the proportion of all outcomes  $(s, r)$  in which the correct indication materializes in a confrontation of two different  $\mathbf{x}$ -vectors. Several pairwise comparisons of this kind were carried out. Typical results are shown in Table 10.4, for

$\text{IncExp}(10 + x_1 + x_2)$ . The upper entry in a table cell shows the success rate in % for  $n = 1,000$ , the lower entry shows that rate for  $n = 2,000$ . Shown in parenthesis is the value of *relbias* for the vectors in question.

“Severe tests” are preferred, that is, confrontations of vectors with a small difference in absolute *relbias*, because the correct decision is then harder to obtain. There is a priori no reason why one of the indicators should always outperform the others in this study. In the five severe tests in Table 10.4,  $H_1$  has, on the whole, better success rates than  $H_2$  and  $H_3$ . The success rate of  $H_1$  improves by doubling the sample size, and tends as expected to be greater when the *relbias* values are further apart. The case  $4G + 8G$  vs.  $8G + 8G$  compares nested  $\mathbf{x}$ -vectors, so it is known beforehand that  $H_2$  and  $H_3$  give perfect success rates.

Table 10.4

Selected pairwise comparisons of auxiliary vectors; percentage of outcomes with correct indication, for the indicators  $H_1$ ,  $H_2$  and  $H_3$ . Within parenthesis, *relbias* in %. Upper entry:  $n = 1,000$  lower entry:  $n = 2,000$ . Response distribution  $\text{IncExp}(10 + x_1 + x_2)$

Cells compared	Percent outcomes with correct indication		
	$H_1$	$H_2$	$H_3$
4G + 8G(0.5) vs.	90.0	100.0	100.0
8G + 8G(0.2)	96.4	100.0	100.0
4G + 2G(1.8) vs.	66.8	86.0	70.7
2G + 8G(1.5)	74.2	89.0	67.4
1G + 8G(4.1) vs.	74.3	70.3	45.0
8G + 1G(3.4)	82.8	78.0	43.3
4G + 1G(4.1) vs.	90.6	61.4	83.9
2G + 2G(3.2)	97.0	68.8	92.3
1G + 2G(7.3) vs.	77.4	77.4	34.5
2G + 1G(6.5)	85.9	85.9	28.8

## 11. Concluding remarks

In this article, we address survey situations where many alternative auxiliary vectors ( $\mathbf{x}$ -vectors) can be created and considered for use in the calibration estimator  $\tilde{Y}_{\text{CAL}}$ . For any given  $\mathbf{x}$ -vector, a certain unknown bias remains in  $\tilde{Y}_{\text{CAL}}$ ; we wish by an appropriate choice of  $\mathbf{x}$ -vector to make that bias as small as possible. Hence we examine the bias ratio defined by (4.2) and (4.3). The component  $\Delta_A$  of the bias ratio was expressed, in (5.8) to (5.10), as product of easily interpreted statistical measures. This led us to suggest several alternative bias indicators, for use in evaluating different  $\mathbf{x}$ -vectors in regard to their capacity to effectively reduce the bias. We studied in particular the indicator  $H_1$  given by (5.12). It functions very well but is geared to a particular study variable  $y$ . However, a typical government survey has many study variables, and for practical reasons it is desirable to use the same  $\mathbf{x}$ -vector in estimating all  $y$ -totals. A compromise becomes necessary. We argued that

the indicator  $H_3$  in (5.12) suits this purpose; it depends on the  $\mathbf{x}_k$  but not on any  $y$ -data. A topic for further research is to develop other indicators (than  $H_3$ ) for the “many  $y$ -variable situation”. Another topic for further work is to examine algorithms for stepwise selection of  $x$ -variables with the indicator  $H_1$ , other than the one used in Section 9.

### Acknowledgements

The authors are grateful to the referees and to the Associate Editor for comments contributing to an improvement of this paper.

### References

- Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie*, I.N.S.E.E., Paris.
- Eltinge, J., and Yansaneh, I. (1997). Diagnostics for the formation of nonresponse adjustment cells with an application to income nonresponse in the US Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.
- Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-98.
- Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.
- Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.
- Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.
- Särndal, C.-E., and Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 4, 251-260.
- Schouten, B. (2007). A selection strategy for weighting variables under a not-missing-at-random assumption. *Journal of Official Statistics*, 23, 51-68.
- Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 101-113.
- Thomsen, I., Kleven, Ø., Wang, J.H. and Zhang, L.C. (2006). Coping with decreasing response rates in Statistics Norway. Recommended practice for reducing the effect of nonresponse. Reports 2006/29. Oslo: Statistics Norway.



# Calibration estimation using exponential tilting in sample surveys

Jae Kwang Kim<sup>1</sup>

## Abstract

We consider the problem of parameter estimation with auxiliary information, where the auxiliary information takes the form of known moments. Calibration estimation is a typical example of using the moment conditions in sample surveys. Given the parametric form of the original distribution of the sample observations, we use the estimated importance sampling of Henmi, Yoshida and Eguchi (2007) to obtain an improved estimator. If we use the normal density to compute the importance weights, the resulting estimator takes the form of the one-step exponential tilting estimator. The proposed exponential tilting estimator is shown to be asymptotically equivalent to the regression estimator, but it avoids extreme weights and has some computational advantages over the empirical likelihood estimator. Variance estimation is also discussed and results from a limited simulation study are presented.

**Key Words:** Benchmarking estimator; Empirical likelihood; Instrumental variable calibration; Importance sampling; Regression estimator.

## 1. Introduction

Consider the problem of estimating  $Y = \sum_{i=1}^N y_i$  for a finite population of size  $N$ . Let  $A$  denote the index set of the sample obtained by a probability sampling scheme. In addition to  $y_i$ , suppose that we also observe a  $p$ -dimensional auxiliary vector  $\mathbf{x}_i$  in the sample such that  $\mathbf{X} = \sum_{i=1}^N \mathbf{x}_i$  is known from an external source. We are interested in estimating  $Y$  using the auxiliary information  $\mathbf{X}$ .

The Horvitz-Thompson (HT) estimator of the form

$$\hat{Y}_d = \sum_{i \in A} d_i y_i, \quad (1)$$

where  $d_i = 1/\pi_i$  is the design weight and  $\pi_i$  is the first order inclusion probability, is unbiased for  $Y$ . But, it does not make use of the information given by  $\mathbf{X}$ . According to Kott (2006), a calibration estimator can be defined as the estimator of the form

$$\hat{Y}_w = \sum_{i \in A} w_i y_i$$

where the weights  $w_i$  satisfy

$$\sum_{i \in A} w_i \mathbf{x}_i = \mathbf{X} \quad (2)$$

and  $\hat{Y}_w$  is asymptotically design unbiased (ADU). Calibration estimation has become very popular in survey sampling because it provides consistency across different surveys and often improves the efficiency. (Särndal 2007).

The regression estimator, using the weights

$$w_i = d_i + (\mathbf{X} - \hat{\mathbf{X}}_d)' \left( \sum_{j \in A} d_j \mathbf{x}_j \mathbf{x}_j' \right)^{-1} d_i \mathbf{x}_i, \quad (3)$$

obtained by minimizing

$$\sum_{i \in A} (w_i - d_i)^2 / d_i$$

subject to constraint (2), is asymptotically design unbiased. Note that if an intercept term is included in the column space of  $\mathbf{X}$  matrix then (2) implies that the population size  $N$  is known. If  $N$  is unknown, one can require that the sum of the final weights are equal to the sum of the design weights. Thus,

$$\sum_{i \in A} w_i = \hat{N}, \quad (4)$$

where

$$\hat{N} = \begin{cases} N & \text{if } N \text{ is known} \\ \sum_{i \in A} d_i & \text{otherwise,} \end{cases}$$

can be imposed as a constraint in addition to (2), which yields the weights

$$w_i = \frac{\hat{N}}{\hat{N}_d} d_i + \left( \mathbf{X} - \frac{\hat{N}}{\hat{N}_d} \hat{\mathbf{X}}_d \right)' \left\{ \sum_{j \in A} d_j (\mathbf{x}_j - \bar{\mathbf{X}}_d) (\mathbf{x}_j - \bar{\mathbf{X}}_d)' \right\}^{-1} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_d), \quad (5)$$

where  $\hat{\mathbf{X}}_d = \sum_{i \in A} d_i \mathbf{x}_i$ ,  $\hat{N}_d = \sum_{i \in A} d_i$ , and  $\bar{\mathbf{X}}_d = \hat{\mathbf{X}}_d / \hat{N}_d$ . We define the regression estimator to be  $\hat{Y}_{\text{reg}} = \sum_{i \in A} w_i y_i$  using the weights (5). The regression estimator can be efficient if  $y_i$  is linearly related with  $\mathbf{x}_i$  (Isaki and Fuller 1982; Fuller 2002), but the weights in the regression estimator can take negative or extremely large values.

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A. E-mail: jkim@iastate.edu.

The empirical likelihood (EL) calibration estimator, discussed by Chen and Qin (1993), Chen and Sitter (1999), Wu and Rao (2006), and Kim (2009), is obtained by maximizing the pseudo empirical likelihood

$$\sum_{i \in A} d_i \ln(w_i)$$

subject to constraints (2) and (4). The solution to the optimization problem can be written as

$$w_i = d_i \frac{1}{\lambda_0 + \lambda_1'(\mathbf{x}_i - \mathbf{X}/\hat{N})}, \quad (6)$$

where  $\lambda_0$  and  $\lambda_1$  satisfy constraints (2), (4), and  $w_i > 0$  for all  $i$ . The EL calibration estimator is asymptotically equivalent to the regression estimator using weights (5) and avoids negative weights if a solution exists, but can result in extremely large weights.

Because the empirical likelihood method requires solving nonlinear equations, the computation can be cumbersome. Furthermore, in some extreme cases,  $\bar{\mathbf{X}} = N^{-1} \sum_{i=1}^N \mathbf{x}_i$  does not belong to the convex hull of the sample  $\mathbf{x}_i$ 's and the solution does not exist. In this extreme situation, the constraint (2) can be relaxed.

Rao and Singh (1997) solved a similar problem by allowing

$$\left| \sum_{i \in A} w_i x_{ij} - X_j \right| \leq \delta_j X_j, \quad j = 1, 2, \dots, p,$$

for some small tolerance level  $\delta_j > 0$  where  $X_j = \sum_{i=1}^N x_{ij}$ . Note that the choice of  $\delta_j = 0$  leads to the exact calibration condition (2). Rao and Singh (1997) chose the tolerance level  $\delta_j$  using a shrinkage factor in the ridge regression but their approach does not directly apply to the empirical likelihood method and the choice of  $\delta_j$  is somewhat unclear. Chambers (1996) and Beaumont and Bocci (2008) also discussed a ridge regression estimation in the context of avoiding extreme weights. Breidt, Claeskens and Opsomer (2005) used penalized spline approach to obtain the ridge calibration. Recently, Park and Fuller (2009) developed a method of obtaining the shrinkage factor  $\delta_j$  using a regression superpopulation model with random components.

Chen, Variyath and Abraham (2008) tackled a similar problem in the context of the empirical likelihood method and proposed a solution by adding an artificial point such that  $\bar{\mathbf{X}} = N^{-1} \sum_{i=1}^N \mathbf{x}_i$  would belong to the convex hull of the augmented  $\mathbf{x}_i$ 's. The proposed estimator in Chen *et al.* (2008) only satisfies the calibration property approximately in the sense that

$$\sum_{i \in A} w_i \mathbf{x}_i - \mathbf{X} = o_p(n^{-1/2}N). \quad (7)$$

This approximate calibration property is attractive because it allows more generality in the choice of weights. In particular, when the dimension of the auxiliary variable  $\mathbf{x}$  is large the calibration constraint (2) can be quite restrictive. As can be seen in Section 2, an estimator satisfying the asymptotic calibration property (7) enjoys most of the desirable properties of the empirical likelihood calibration estimator and is computationally efficient.

In this paper, we consider a class of empirical-likelihood-type estimators that satisfy the approximate calibration property (7). In Section 2, the idea of estimated importance sampling of Henmi *et al.* (2007) is discussed and a new estimator using this methodology is proposed. In Section 3, a weight trimming technique to avoid extreme calibration weights is proposed. In Section 4, variance estimation of the proposed estimator is discussed. In Section 5, results from a simulation study are presented. Concluding remarks are made in Section 6.

## 2. Proposed method

To introduce the proposed method, we first discuss estimated importance sampling introduced by Henmi *et al.* (2007). Suppose that  $\mathbf{x}_i$  is observed throughout the population but  $y_i$  is observed only in the sample. We assume a superpopulation model for  $\mathbf{x}_i$  with density  $f(\mathbf{x}; \boldsymbol{\eta})$  known up to a parameter  $\boldsymbol{\eta} \in \Omega$ . The superpopulation model characterized by the density  $f(\mathbf{x}; \boldsymbol{\eta})$  is a working model in the sense that the model is used to derive a model-assisted estimator (Särndal, Swenson and Wretman 1992).

Let  $\hat{\boldsymbol{\eta}}$  be the pseudo maximum likelihood estimator of  $\boldsymbol{\eta}$  computed from the sample

$$\hat{\boldsymbol{\eta}} = \arg \max_{\boldsymbol{\eta} \in \Omega} \sum_{i \in A} d_i \ln \{f(\mathbf{x}_i; \boldsymbol{\eta})\}$$

and let  $\boldsymbol{\eta}_{0,N}$  be the maximum likelihood estimator of  $\boldsymbol{\eta}$  computed from the population

$$\boldsymbol{\eta}_{0,N} = \arg \max_{\boldsymbol{\eta} \in \Omega} \sum_{i=1}^N \ln \{f(\mathbf{x}_i; \boldsymbol{\eta})\}.$$

Following Henmi *et al.* (2007), we can construct the following estimated importance weight

$$w_i = d_i \frac{f(\mathbf{x}_i; \boldsymbol{\eta}_{0,N})}{f(\mathbf{x}_i; \hat{\boldsymbol{\eta}})}. \quad (8)$$

To discuss the asymptotic properties of the estimator using the weights in (8), assume a sequence of the finite populations and the samples, as in Isaki and Fuller (1982), such that

$$\sum_{i \in A} d_i (\mathbf{x}'_i, y_i)' (\mathbf{x}'_i, y_i) - \sum_{i=1}^N (\mathbf{x}'_i, y_i)' (\mathbf{x}'_i, y_i) = O_p(n^{-1/2}N)$$

for all possible  $A$  and for each  $N$ . The following theorem presents some asymptotic properties of the estimator with the estimated importance weights in (8).

*Theorem 1. Under the regularity conditions given in Appendix A, the estimator  $\hat{Y}_w = \sum_{i \in A} w_i y_i$ , with the  $w_i$  defined by (8), satisfies*

$$\sqrt{n}N^{-1}(\hat{Y}_w - \hat{Y}_l) = o_p(1), \quad (9)$$

where

$$\hat{Y}_l = \hat{Y}_d - \hat{\Sigma}'_{sy} \hat{\Sigma}_{ss}^{-1} \hat{S}_{0d}, \quad (10)$$

$\hat{Y}_d$  is defined in (1),  $\hat{S}_{0d} = \sum_{i \in A} d_i s_{i0}$ ,  $\hat{\Sigma}_{sy} = N^{-1} \sum_{i \in A} d_i s_{i0} y_i$ , and  $\hat{\Sigma}_{ss} = N^{-1} \sum_{i \in A} d_i s_{i0}^{\otimes 2}$ . Here,  $s_{i0} = \partial \ln f(\mathbf{x}_i; \boldsymbol{\eta}) / \partial \boldsymbol{\eta}|_{\boldsymbol{\eta}=\boldsymbol{\eta}_{0,N}}$  and the notation  $B^{\otimes 2}$  denotes  $BB'$ .

The proof of Theorem 1 is presented in Appendix A. Because  $\mathbf{S}_{0N} \equiv \sum_{i=1}^N \mathbf{s}_{i0} = \mathbf{0}$ , we can write (10) as

$$\hat{Y}_l = \hat{Y}_d + \hat{\Sigma}'_{sy} \hat{\Sigma}_{ss}^{-1} (\mathbf{S}_{0N} - \hat{S}_{0d}),$$

which is a regression estimator of  $Y$  using  $\mathbf{s}_i(\boldsymbol{\eta}_{0,N})$  as the auxiliary variable. Therefore, under regularity conditions, the proposed estimator using estimated importance sampling is asymptotically unbiased and has asymptotic variance no greater than that of the direct estimator  $\hat{Y}_d$ . Note that the validity of Theorem 1 does not require that the working model  $f(\mathbf{x}; \boldsymbol{\eta})$  be true.

If the density of  $\mathbf{x}_i$  is a multivariate normal density, then the weights in (8) become

$$w_i = d_i \frac{\phi(\mathbf{x}_i; \bar{\mathbf{X}}_N, \boldsymbol{\Sigma}_{xx,N})}{\phi(\mathbf{x}_i; \bar{\mathbf{X}}_d, \hat{\boldsymbol{\Sigma}}_{xx,d})}, \quad (11)$$

where  $\bar{\mathbf{X}}_d$  is defined after (5),  $\hat{\boldsymbol{\Sigma}}_{xx,d} = \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_d)^{\otimes 2} / \hat{N}_d$ ,  $\boldsymbol{\Sigma}_{xx,N} = \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{X}}_N)^{\otimes 2} / N$ , and  $\phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  is the density of the multivariate normal distribution with mean  $\boldsymbol{\mu}$  and variance-covariance matrix  $\boldsymbol{\Sigma}$ . If  $\boldsymbol{\Sigma}_{xx,N}$  is unknown and only  $\bar{\mathbf{X}}_N$  is available, then we can use

$$w_i = d_i \frac{\phi(\mathbf{x}_i; \bar{\mathbf{X}}_N, \hat{\boldsymbol{\Sigma}}_{xx,d})}{\phi(\mathbf{x}_i; \bar{\mathbf{X}}_d, \hat{\boldsymbol{\Sigma}}_{xx,d})}. \quad (12)$$

Tillé (1998) derived weights similar to those in (12) in the context of conditional inclusion probabilities.

In general, the parametric model for  $\mathbf{x}_i$  is unknown. Thus, we consider an approximation for the importance weights in (8) using the Kullback-Leibler information criterion for distance. Let  $f(\mathbf{x})$  be a given density for  $\mathbf{x}$  and let  $P_0$  be the set of densities that satisfy the calibration constraint. That is,

$$P_0 = \left\{ f_0(\mathbf{x}); \int f_0(\mathbf{x}) d\mathbf{x} = 1, \int \mathbf{x} f_0(\mathbf{x}) d\mathbf{x} = \bar{\mathbf{X}}_N \right\}.$$

The optimization problem using Kullback-Leibler distance can be expressed as

$$\min_{f_0 \in P_0} \int f_0(\mathbf{x}) \ln \left\{ \frac{f_0(\mathbf{x})}{f(\mathbf{x})} \right\} d\mathbf{x}. \quad (13)$$

The solution to (13) is

$$f_0(\mathbf{x}) = f(\mathbf{x}) \frac{\exp(\hat{\boldsymbol{\lambda}}' \mathbf{x})}{E\{\exp(\hat{\boldsymbol{\lambda}}' \mathbf{x})\}} \quad (14)$$

where  $\hat{\boldsymbol{\lambda}}$  satisfies  $\int \mathbf{x} f_0(\mathbf{x}) d\mathbf{x} = \bar{\mathbf{X}}_N$ . Thus, the estimated importance weights in (8) using the optimal density in (14) can be written

$$w_i = d_i \frac{f_0(\mathbf{x}_i)}{f(\mathbf{x}_i)} = d_i \exp(\hat{\boldsymbol{\lambda}}_0 + \hat{\boldsymbol{\lambda}}_1' \mathbf{x}_i) \quad (15)$$

where  $\hat{\boldsymbol{\lambda}}_0$  and  $\hat{\boldsymbol{\lambda}}_1$  satisfy constraint (2) and (4). The shift from  $f(\mathbf{x})$  to  $f_0(\mathbf{x})$  in (14) is called exponential tilting. Thus, an estimator using the weight (15) satisfying the calibration constraints (2) and (4) can be called an exponential tilting (ET) calibration estimator. That is, we define the ET calibration estimator as

$$\hat{Y}_{ET} = \sum_{i \in A} d_i \exp(\hat{\boldsymbol{\lambda}}_0 + \hat{\boldsymbol{\lambda}}_1' \mathbf{x}_i) y_i, \quad (16)$$

where  $\hat{\boldsymbol{\lambda}}_0$  and  $\hat{\boldsymbol{\lambda}}_1$  satisfy constraint (2) and (4). Estimators based on exponential tilting have been used in various contexts. For examples, see Efron (1981), Kitamura and Stutzer (1997), and Imbens (2002). When  $N$  is known, Folsom (1991) and Deville, Särndal and Sautory (1993) developed the estimator (16) using a very different approach.

To compute  $\lambda_0$  and  $\lambda_1$  in (16), because of the calibration constraints (2) and (4), we need to solve the following estimating equations:

$$\hat{U}_0(\boldsymbol{\lambda}) \equiv \sum_{i \in A} d_i \exp(\lambda_0 + \lambda_1' \mathbf{x}_i) - \hat{N} = 0 \quad (17)$$

$$\hat{U}_1(\boldsymbol{\lambda}) \equiv \sum_{i \in A} d_i \exp(\lambda_0 + \lambda_1' \mathbf{x}_i) \mathbf{x}_i - \mathbf{X} = \mathbf{0}, \quad (18)$$

where  $\boldsymbol{\lambda}' = (\lambda_0, \lambda_1')$ . Writing  $\hat{\mathbf{U}}' = (\hat{U}_0, \hat{U}_1')$ , we can use the Newton-type algorithm of the form

$$\hat{\boldsymbol{\lambda}}_{(t+1)} = \hat{\boldsymbol{\lambda}}_{(t)} - \left\{ \frac{\partial}{\partial \boldsymbol{\lambda}'} \hat{\mathbf{U}}(\hat{\boldsymbol{\lambda}}_{(t)}) \right\}^{-1} \hat{\mathbf{U}}(\hat{\boldsymbol{\lambda}}_{(t)})$$

and the solution can be written

$$\hat{\boldsymbol{\lambda}}_{1(t+1)} = \hat{\boldsymbol{\lambda}}_{1(t)} + \left\{ \sum_{i \in A} w_{i(t)} (\mathbf{x}_i - \bar{\mathbf{X}}_{w(t)})^{\otimes 2} \right\}^{-1} \left( \mathbf{X} - \sum_{i \in A} w_{i(t)} \mathbf{x}_i \right), \quad (19)$$



where  $w_{i(t)} = d_i \exp(\hat{\lambda}_{0(t)} + \hat{\lambda}'_{1(t)} \mathbf{x}_i)$  and  $\bar{\mathbf{X}}_{w(t)} = \sum_{i \in A} w_{i(t)} \mathbf{x}_i / \sum_{i \in A} w_{i(t)}$ , with the initial values  $\hat{\lambda}_{1(0)} = \mathbf{0}$ . Once  $\hat{\lambda}_{1(t)}$  is computed by (19),  $\hat{\lambda}_{0(t)}$  is computed by

$$\exp(\hat{\lambda}_{0(t)}) = \frac{\hat{N}}{\sum_{i \in A} d_i \exp(\hat{\lambda}'_{1(t)} \mathbf{x}_i)}. \quad (20)$$

Note that,  $w_{i(0)} = d_i \hat{N} / \hat{N}_d$  since  $\hat{\lambda}_{1(0)} = \mathbf{0}$ . Because  $\hat{\mathbf{U}}(\boldsymbol{\lambda})$  is twice continuously differentiable and convex in  $\boldsymbol{\lambda}$ , the sequence  $\hat{\lambda}_{1(t)}$  always converges if the solution to  $\hat{\mathbf{U}}(\boldsymbol{\lambda}) = \mathbf{0}$  exists (Givens and Hoeting 2005). The convergence rate is quadratic in the sense that

$$|\hat{\lambda}_{1(t+1)} - \hat{\lambda}_1| \leq C |\hat{\lambda}_{1(t)} - \hat{\lambda}_1|^2$$

for some constant  $C$ , where  $\hat{\lambda}_1 = \lim_{t \rightarrow \infty} \hat{\lambda}_{1(t)}$ .

By construction, the  $t$ -step exponential tilting (ET) estimator, defined by

$$\hat{Y}_{\text{ET}(t)} = \sum_{i \in A} d_i \exp(\hat{\lambda}_{0(t)} + \hat{\lambda}'_{1(t)} \mathbf{x}_i) y_i \quad (21)$$

where  $\hat{\lambda}_{0(t)}$  and  $\hat{\lambda}_{1(t)}$  are computed by (19) and (20), satisfies the calibration constraint (2) for sufficiently large  $t$ . By the recursive form in (19) with  $\hat{\lambda}_{1(0)} = \mathbf{0}$ , we can write

$$\hat{\lambda}_{1(t)} = \sum_{j=0}^{t-1} (\mathbf{S}_{xx, w(j)})^{-1} (\tilde{\mathbf{X}}_N - \bar{\mathbf{X}}_{w(j)}), \quad (22)$$

where  $\tilde{\mathbf{X}}_N = \mathbf{X} / \hat{N}$  and  $\mathbf{S}_{xx, w(j)} = \sum_{i \in A} w_{i(t)} (\mathbf{x}_i - \bar{\mathbf{X}}_{w(t)})^{\otimes 2} / \hat{N}$ . Thus, the  $t$ -step ET estimator (21) can be written as

$$\hat{Y}_{\text{ET}(t)} = \hat{N} \frac{\sum_{i \in A} d_i g_{i(t)} y_i}{\sum_{i \in A} d_i g_{i(t)}},$$

where

$$g_{i(t)} = \prod_{j=0}^{t-1} \frac{\phi(\mathbf{x}_i; \tilde{\mathbf{X}}_N, \mathbf{S}_{xx, w(j)})}{\phi(\mathbf{x}_i; \bar{\mathbf{X}}_{w(j)}, \mathbf{S}_{xx, w(j)})}.$$

The following theorem presents some asymptotic properties of the exponential tilting estimator.

*Theorem 2. The  $t$ -step ET estimator (21) based on equations (19) and (20) satisfies*

$$\sqrt{n} N^{-1} (\hat{Y}_{\text{ET}(t)} - \hat{Y}_{\text{reg}}) = o_p(1), \quad (23)$$

for each  $t = 1, 2, \dots$ , where  $\hat{Y}_{\text{reg}}$  is the regression estimator using the regression weight in (5).

The proof of Theorem 2 is presented in Appendix B. Theorem 2 presents the asymptotic equivalence between the  $t$ -step ET estimator and the regression estimator. Unlike the regression estimator, the weights of the ET estimator are always positive. For sufficiently large  $t$ , the  $t$ -step ET estimator satisfies the calibration constraint (2). Deville and Särndal (1992) proved the result (23) for the special case of  $t \rightarrow \infty$ .

*Remark 1. The one-step ET estimator, defined by  $\hat{Y}_{\text{ET}(1)}$ , has a closed-form tilting parameter*

$$\hat{\lambda}_{1(1)} = \left\{ \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_d)^{\otimes 2} / \hat{N}_d \right\}^{-1} (\tilde{\mathbf{X}}_N - \bar{\mathbf{X}}_d), \quad (24)$$

where  $\tilde{\mathbf{X}}_N = \mathbf{X} / \hat{N}$  and  $\bar{\mathbf{X}}_d = \sum_{i \in A} d_i \mathbf{x}_i / \sum_{i \in A} d_i$ . By Theorem 2, the one-step ET estimator is asymptotically equivalent to the regression estimator, but the calibration constraint (2) is not necessarily satisfied. Using Theorem 2 applied to  $\mathbf{x}_i$  instead of  $y_i$ , the one-step ET estimator can be shown to satisfy the approximate calibration constraint described in (7).

*Remark 2. The ET estimator can also be derived by finding the weights that minimize*

$$Q(w) = \sum_{i \in A} w_i \ln \left( \frac{w_i}{d_i} \right) \quad (25)$$

subject to constraints (2) and (4). The objective function (25) is often called the minimum discrimination function. The minimum value of  $Q(w)$  is zero if (4) is the only calibration constraint and is monotonically increasing if additional calibration constraints are imposed.

### 3. Instrumental-variable calibration

We consider some extension of the proposed method in Section 2 to a more general class of ET calibration estimator using instrumental-variables. Use of instrumental-variable in the calibration estimation has been discussed in Estevao and Särndal (2000) and Kott (2003) in some limited simulations. Let  $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i)$  be an instrumental-variable derived from  $\mathbf{x}_i$ , where the function  $\mathbf{z}(\cdot)$  is to be determined. The instrumental-variable exponential tilting (IVET) estimator using the instrumental variable  $\mathbf{z}_i$  can be defined as

$$\hat{Y}_{\text{IVET}} = \sum_{i \in A} w_i y_i = \sum_{i \in A} d_i \exp(\hat{\lambda}_0 + \hat{\lambda}'_1 \mathbf{z}_i) y_i, \quad (26)$$

where  $\hat{\lambda}_0$  and  $\hat{\lambda}_1$  are computed from (2) and (4). Note that the IVET estimator (26) is a class of estimators indexed by  $\mathbf{z}_i$ . The instrumental-variable approach defined in (26) provides more flexibility in creating the ET estimator. The choice of  $\mathbf{z}_i = \mathbf{x}_i$  leads to the standard ET estimator in (16) but some transformation  $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i)$  can make the resulting ET estimator in (26) more attractive in practice. The solution to the calibration equations can be obtained iteratively by

$$\hat{\lambda}_{1(t+1)} = \hat{\lambda}_{1(t)} + \left\{ \sum_{i \in A} w_{i(t)} (\mathbf{x}_i - \bar{\mathbf{X}}_{w(t)}) (\mathbf{z}_i - \bar{\mathbf{Z}}_{w(t)})' \right\}^{-1} \left( \mathbf{X} - \sum_{i \in A} w_{i(t)} \mathbf{x}_i \right), \quad (27)$$

where  $w_{i(t)} = d_i \exp(\hat{\lambda}_{0(t)} + \hat{\lambda}'_{1(t)} \mathbf{z}_i)$  and  $\bar{\mathbf{Z}}_{w(t)} = \sum_{i \in A} w_{i(t)} \mathbf{z}_i / \sum_{i \in A} w_{i(t)}$ , with equation (20) unchanged and  $\hat{\lambda}_{1(0)} = \mathbf{0}$ .

The IVET estimator (26) is useful in creating the final weights that have less extreme values. Since the final weight in (26) is a function of  $\mathbf{z}_i$ , we can make  $g_i = w_i/d_i$  bounded by making  $\mathbf{z}_i$  bounded. To create bounded  $\mathbf{z}_i$ , we can use a trimmed version of  $\mathbf{x}_i$ , noted by  $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{ip})$ , where

$$z_{ij} = \begin{cases} x_{ij} & \text{if } |x_{ij} - \bar{x}_j| \leq C_j S_j \\ \bar{x}_j + C_j S_j & \text{if } x_{ij} > \bar{x}_j + C_j S_j \\ \bar{x}_j - C_j S_j & \text{if } x_{ij} < \bar{x}_j - C_j S_j, \end{cases} \quad (28)$$

$\bar{x}_j = N^{-1} \sum_{i \in A} d_i x_{ij}$ ,  $S_j^2 = N^{-1} \sum_{i \in A} d_i (x_{ij} - \bar{x}_j)^2$ , and  $C_j$  is a threshold for detecting outliers, for example,  $C_j = 3$ . Thus, the IVET estimator using the instrumental-variable obtained by trimming  $\mathbf{x}_i$  can be used as an alternative approach to weight trimming.

Instead of using the trimmed instrumental variable  $\mathbf{z}_i$  in (28), we can consider the following instrumental variable

$$\mathbf{z}_i = \mathbf{x}_i \Phi_i$$

for some symmetric matrix  $\Phi_i$  such that  $\mathbf{z}_i$  is bounded. Some suitable choice of  $\Phi_i$  can also improve the efficiency of the resulting IVET estimator. To see this, using the same argument from Theorem 2, the instrumental-variable ET estimator (26) using equations (20) and (27) is asymptotically equivalent to

$$\hat{Y}_{IV, \text{reg}} = \tilde{Y}_d + (\mathbf{X} - \tilde{\mathbf{X}}_d)' \hat{\mathbf{B}}_z \quad (29)$$

where

$$(\tilde{\mathbf{X}}_d', \tilde{Y}_d') = \left( \frac{\hat{N}}{\hat{N}_d} \right) (\hat{\mathbf{X}}_d', \hat{Y}_d')$$

and

$$\hat{\mathbf{B}}_z = \left\{ \sum_{i \in A} d_i (\mathbf{z}_i - \bar{\mathbf{z}}_d) (\mathbf{x}_i - \bar{\mathbf{x}}_d)' \right\}^{-1} \sum_{i \in A} d_i (\mathbf{z}_i - \bar{\mathbf{z}}_d) y_i. \quad (30)$$

The estimator (29) takes the form of a regression estimator and is called the instrumental-variable regression estimator. Thus, under the choice of  $\mathbf{z}_i = \Phi_i \mathbf{x}_i$ , the instrumental-variable regression estimator can be written as (29) with

$$\hat{\mathbf{B}}_z = \left\{ \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{x}}_d) \Phi_i (\mathbf{x}_i - \bar{\mathbf{x}}_d)' \right\}^{-1} \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{x}}_d) \Phi_i y_i$$

and its variance is minimized for  $\Phi_i = V_i^{-1}$  where  $V_i$  is the model-variance of  $y_i$  given  $\mathbf{x}_i$  (Fuller 2009). The model-variance is the variance under the working superpopulation model for the regression of  $y_i$  on  $\mathbf{x}_i$ . Thus, instrumental-variable can be used to improve the efficiency of the resulting calibration estimator, in addition to avoid extreme final weights. Furthermore, the optimal instrumental-variable can be trimmed as in (28) to make the final weights bounded. Further investigation of the optimal choice of  $\Phi$  is beyond the scope of this paper and will be a topic of future research.

*Remark 3.* Deville and Särndal (1992) also considered range-restricted calibration weights of the form

$$w_i = d_i g_i(\hat{\lambda}) = d_i \frac{L(U-1) + U(1-L) \exp(K \hat{\lambda}' \mathbf{x}_i)}{(U-1) + (1-L) \exp(K \hat{\lambda}' \mathbf{x}_i)}, \quad (31)$$

where  $K = (U-L)/\{(1-L)(U-1)\}$ , for some  $L$  and  $U$  such that  $0 < L < 1 < U$ . If calibration constraints (2) and (4) are to be satisfied, then we can use  $\hat{\lambda}_0 + \hat{\lambda}'_1 \mathbf{x}_i$  instead of  $\hat{\lambda}'_1 \mathbf{x}_i$  in (31). The resulting calibration estimator is asymptotically equivalent to the regression estimator using the weights in (5) while the IVET estimator is asymptotically equivalent to the instrumental-variable regression estimator (29). Computation for obtaining  $\hat{\lambda}$  is somewhat complicated because  $\partial g_i(\lambda)/\partial \lambda$  is not easy to evaluate in (31). In the IVET estimator, the computation, given by (27), is straightforward.

To compare the proposed weight with existing methods, we consider an artificial example of a simple random sample with size  $n=5$  where  $x_k = k$ ,  $k=1, 2, \dots, 5$ . Calculations are for three population means of  $x$ ;  $\bar{X}_N = 3$ ,  $\bar{X}_N = 4.5$ , and  $\bar{X}_N = 6$ . Table 1 presents the resulting weights for the regression estimator, the empirical likelihood (EL) estimator, the  $t$ -step ET estimator (16) with  $t=1$  and  $t=10$ , and the  $t$ -step instrumental variable exponential tilting (IVET) estimator (26) with  $t=1$  and  $t=10$ . For the IVET estimator, the instrumental variable  $z_i$  is created by

$$z_i = \begin{cases} 1.5 & \text{if } x_i \leq 1.5 \\ x_i & \text{if } x_i \in (1.5, 4.5) \\ 4.5 & \text{if } x_i \geq 4.5. \end{cases}$$

The last column of Table 1 presents the estimated mean of  $X$  using the respective calibration weights. All the weights are equal to  $1/n=0.2$  for  $\bar{X}_N = 3$ . The regression estimator is linearly increasing in  $x_i$  but has negative weights for the population with  $\bar{X}_N = 4.5$  and  $\bar{X}_N = 6$ . For the population where  $\bar{X}_N = 6$ , the weights could not be computed for the EL method because  $\bar{X}_N$  is outside the range of the sample  $x_i$ 's. In this extreme case of  $\bar{X}_N = 6$ , the ET method provides nonnegative weights by sacrificing the calibration constraint and the EL estimator has more extreme weights than the ET estimator or IVET estimator in the sense that the weight for  $k=5$  is the largest among the estimators considered. The weight for the one-step ET estimator is close to that of the regression estimator for large  $x_i$  but it is close to that of EL estimator for small  $x_i$ . The 10-step ET estimators has better calibration properties in the sense of smaller value of squared error,  $(\sum_{k=1}^5 w_k x_k - \bar{X}_N)^2$ , than the one-step ET estimator. The ET estimator and the IVET estimator provide almost the same estimates of  $\bar{X}_N$  for both  $t$ , but the IVET estimator produces less extreme weights than the ET estimator.

**Table 1**  
An example of calibration weights with a sample of size  $n = 5$

Method	$\bar{X}_N$	$x_i$					$\hat{X}_N$
		1	2	3	4	5	
Reg.	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	-0.100	0.050	0.200	0.035	0.500	4.5
	6.0	-0.400	-0.100	0.200	0.500	0.800	6.0
EL	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	0.033	0.043	0.063	0.115	0.746	4.5
	6.0	N/A	N/A	N/A	N/A	N/A	N/A
ET ( $t = 1$ )	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	0.027	0.057	0.100	0.255	0.540	4.2
	6.0	0.002	0.009	0.039	0.173	0.777	4.7
ET ( $t = 10$ )	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	0.009	0.027	0.078	0.227	0.659	4.5
	6.0	0.000	0.000	0.000	0.001	0.999	5.0
IVET ( $t = 1$ )	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	0.030	0.047	0.121	0.309	0.493	4.2
	6.0	0.003	0.006	0.041	0.267	0.683	4.6
IVET ( $t = 10$ )	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	0.007	0.015	0.066	0.294	0.618	4.5
	6.0	0.000	0.000	0.000	0.087	0.913	4.9

Reg., Regression estimator; EL, empirical likelihood; ET, exponential tilting; IVET, instrumental variable exponential tilting; N/A, Not applicable.

#### 4. Variance estimation

We now discuss variance estimation of the ET calibration estimators of Sections 2 and 3. Because the estimated parameter  $(\hat{\lambda}_0, \hat{\lambda}'_1)$  in the ET calibration estimator (16) has some sampling variability, variance estimation method should take into account of this sampling variability of these estimated parameters. In this case, variance estimation can be often obtained by a linearization method or by a replication method (Wolter 2007). For the discussion of the linearization method, let the variance of the HT estimator (1) be consistently estimated by

$$\hat{V}(\hat{Y}_d) = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} y_i y_j. \quad (32)$$

The linearization variance estimator for the ET estimator can be obtained by the linearization variance formula for the regression estimator, as in Deville and Särndal (1992), using the asymptotic equivalence between the ET calibration estimator and the regression estimator, as shown in Theorem 2. Specifically, if the population size  $N$  is known, a linearization variance estimator of the IVET estimator in (26) can be written as

$$\hat{V}(\hat{Y}_{IVET}) = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} g_i g_j \hat{e}_i \hat{e}_j \quad (33)$$

where  $\Omega_{ij}$  are the coefficients of the variance estimator in (32),  $g_i = w_i/d_i$  is the weight adjustment factor, and  $\hat{e}_i = y_i - \bar{Y}_d - (\mathbf{x}_i - \bar{\mathbf{X}}_d)' \hat{\mathbf{B}}_z$ , where  $\hat{\mathbf{B}}_z$  is defined in (30). The choice of  $\mathbf{z}_i = \mathbf{x}_i$  in (33) gives the linearized variance estimator for the ET estimator in (16). Consistency of the variance estimator (33) can be found in Kim and Park (2010).

For the one-step ET estimator, a replication method can be easily implemented. Let the replication variance estimator be of the form

$$\hat{V}_{\text{rep}} = \sum_{k=1}^L c_k (\hat{Y}_d^{(k)} - \hat{Y}_d)^2, \quad (34)$$

where  $L$  is the number of replication,  $c_k$  is the replication factor associated with replicate  $k$ ,  $\hat{Y}_d^{(k)} = \sum_{i \in A} d_i^{(k)} y_i$ , and  $d_i^{(k)}$  is the  $k^{\text{th}}$  replicate of the design weight  $d_i$ . For example, the replication variance estimator (34) includes the jackknife and the bootstrap (see Rust and Rao 1996). Assume that the replication variance estimator (34) is a consistent estimator for the variance of  $\hat{Y}_d$ . The  $k^{\text{th}}$  replicate of the one-step ET estimator can be computed by

$$\hat{Y}_{ET(1)}^{(k)} = \sum_{i \in A} d_i^{(k)} \exp(\hat{\lambda}_{0(1)}^{(k)} + \hat{\lambda}_{1(1)}^{(k)} \mathbf{z}_i) y_i \quad (35)$$

where

$$\hat{\lambda}_{1(1)}^{(k)} = \left\{ \sum_{i \in A} d_i^{(k)} (\mathbf{x}_i - \bar{\mathbf{X}}_d^{(k)}) (\mathbf{z}_i - \bar{\mathbf{Z}}_d^{(k)})' / \hat{N}_d^{(k)} \right\}^{-1} (\mathbf{X} / \hat{N}^{(k)} - \bar{\mathbf{X}}_d^{(k)}),$$

$$\hat{N}^{(k)} = \begin{cases} N & \text{if } \hat{N} = N \\ \hat{N}_d^{(k)} = \sum_{i \in A} d_i^{(k)} & \text{if } \hat{N} = \hat{N}_d, \end{cases}$$

$$(\bar{\mathbf{X}}_d^{(k)}, \bar{\mathbf{Z}}_d^{(k)}) = \frac{\sum_{i \in A} d_i^{(k)} (\mathbf{x}_i, \mathbf{z}_i)}{\sum_{i \in A} d_i^{(k)}},$$

and

$$\exp(\hat{\lambda}_{0(1)}^{(k)}) = \frac{\hat{N}}{\sum_{i \in A} d_i^{(k)} \exp(\mathbf{z}_i' \hat{\lambda}_{1(1)}^{(k)})}.$$

The replication variance estimator defined by



$$\hat{V}_{\text{rep}} = \sum_{k=1}^L c_k (\hat{Y}_{\text{ET}}^{(k)} - \hat{Y}_{\text{ET}})^2, \quad (36)$$

where  $\hat{Y}_{\text{ET}}^{(k)}$  is defined in (35), can be used to estimate the variance of the ET calibration estimator in (26).

## 5. Simulation study

To study the finite sample performance of the proposed estimators, we performed a limited simulation study. In the simulation, two finite populations of size  $N = 10,000$  were independently generated. In population A, the finite population is generated from an infinite population specified by  $x_i \sim \exp(1) + 1$ ;  $y_i = 3 + x_i + x_i e_i$ ,  $e_i | x_i \sim N(0, 1)$ ;  $z_i | (x_i, y_i) \sim \chi^2(1) + |y_i|$ . In population B,  $(x_i, e_i, z_i)$  are the same as in population A but  $y_i = (5 - 1/\sqrt{8}) + 1/\sqrt{8}(x_i - 2)^2 + e_i$ . The auxiliary variable,  $x_i$ , is used for calibration and  $z_i$  is the measure of size used for unequal probability sampling. From both of the finite populations generated,  $M = 10,000$  Monte Carlo samples of size  $n$  were independently generated under two sampling schemes described below. The parameter of interest is the population mean of  $y$  and we assume that the population size  $N$  is known.

The simulation setup can be described as a  $2 \times 2 \times 8 \times 2$  factorial design with four factors. The factors are (a) two types of finite populations, (b) Sampling mechanism: simple random sampling and probability proportional to size ( $z_i$ ) sampling with replacement, (c) Calibration method: no calibration, the regression estimator, the EL method in (6) with  $t=1$  and  $t=10$ , the  $t$ -step ET method in (21) with  $t=1$  and  $t=10$ , and the IVET method (26) with  $t=1$  and  $t=10$ , (d) sample size:  $n=100$  and  $n=200$ . Since  $N$  is assumed to be known, the calibration estimators are computed to satisfy  $\sum_{i=1}^n w_i(1, x_i) = (1, \bar{X}_N)$  in both populations. For the IVET method (26), the instrumental variable  $z_i$  is created using the definitions in (28) with threshold  $C=3$ .

Using the Monte Carlo samples generated as above, the biases and the mean squared errors of the eight estimators of the population mean of  $y$ , the variable of interest, were computed and are presented in Table 2. The calibration estimators are biased but the bias is small if the regression model holds or the sample size is large. In population A, the linear regression model holds and the regression estimator is efficient in terms of mean squared errors. However, the regression estimator is not efficient in population B because the model used for the regression estimator is not a good fit. The seven calibration estimators show similar performances for the larger sample size. The 10-step IVET estimator performs as well as the regression estimator in population A, and it shows slightly better performance than the other

six calibration estimators. In population B, the 10-step IVET estimator performs the best among the calibration estimators considered.

In addition to point estimation, variance estimation was also considered. We considered only the variance estimation for the  $t$ -step ET estimators and IVET estimators. The linearization variance estimator in (33) and the replication variance estimator in (36) were computed for each estimator in each sample. In the replication method, the jackknife method was used by deleting one element for each replication. The relative biases of the variance estimators were computed by dividing the Monte Carlo bias of the variance estimator by the Monte Carlo variance. The Monte Carlo relative biases of the linearization variance estimators and the replication variance estimators are presented in Table 3. The theoretical relative bias of the variance estimators is of order  $o(1)$ , which is consistent with the simulation results in Table 3. The linearization variance estimator slightly underestimates the true variance because it ignores the second order term in the Taylor linearization. The replication variance estimator shows slight positive bias in the simulation. The biases of the variance estimators are generally smaller in absolute values in population A because the linear model holds. In population B, variance estimators for the IVET estimator are less biased than those for the ET estimator because of less extreme weights used by the IVET estimator.

## 6. Concluding remarks

We have considered the problem of estimating  $Y$  with auxiliary information of the form  $E\{U(\mathbf{X})\} = 0$  with some known function  $U(\cdot)$ . The class of the linear estimators of the form  $\hat{Y} = \sum_{i \in A} w_i y_i$  with  $\sum_{i \in A} w_i \{1, U(\mathbf{x}_i)\} = (\hat{N}, 0)$  and  $w_i > 0$  is considered. If the density  $f(\mathbf{x}; \boldsymbol{\eta})$  of  $X$  is known up to  $\boldsymbol{\eta} \in \Omega$ , then an efficient estimation can be implemented using the estimated importance weight

$$w_i \propto d_i \frac{f(\mathbf{x}_i; \boldsymbol{\eta}_{0,N})}{f(\mathbf{x}_i; \hat{\boldsymbol{\eta}})},$$

where  $d_i$  are the initial weights and where  $\boldsymbol{\eta}_{0,N}$  and  $\hat{\boldsymbol{\eta}}$  are the maximum likelihood estimators of  $\boldsymbol{\eta}$  based on the population and the sample, respectively. If the parametric form of  $f(\mathbf{x}; \boldsymbol{\eta})$  is unknown, then the exponential tilting weights of the form

$$w_{i(\lambda)} \propto \exp\{\boldsymbol{\lambda}' U(\mathbf{x}_i)\}$$

can be used, where  $\boldsymbol{\lambda}$  is determined to satisfy

$$\sum_{i \in A} w_{i(\lambda)} U(\mathbf{x}_i) = 0. \quad (37)$$

Table 2

Monte Carlo Biases and Monte Carlo Mean squared errors of the point estimators for the mean of  $y$ , based on 10,000 Monte Carlo samples

Population	Sample Size	Estimator	SRS		PPS	
			Bias	MSE	Bias	MSE
A	100	No Calibration	0.00	0.02398	0.00	0.02023
		Regression estimator	0.00	0.01261	0.00	0.01289
		EL estimator ( $t = 1$ )	0.01	0.01369	0.01	0.01353
		EL estimator ( $t = 10$ )	0.00	0.01285	0.00	0.01289
		ET estimator ( $t = 1$ )	0.01	0.01334	0.01	0.01353
		ET estimator ( $t = 10$ )	0.00	0.01269	0.00	0.01289
		IVET estimator ( $t = 1$ )	0.01	0.01309	0.01	0.01330
		IVET estimator ( $t = 10$ )	0.00	0.01263	0.00	0.01289
	200	No Calibration	0.00	0.01069	0.00	0.00925
		Regression estimator	0.00	0.00595	0.00	0.00568
		EL estimator ( $t = 1$ )	0.01	0.00632	0.01	0.00604
		EL estimator ( $t = 10$ )	0.00	0.00597	0.00	0.00568
		ET estimator ( $t = 1$ )	0.00	0.00616	0.01	0.00578
		ET estimator ( $t = 10$ )	0.00	0.00596	0.00	0.00568
		IVET estimator ( $t = 1$ )	0.00	0.00605	0.01	0.00574
		IVET estimator ( $t = 10$ )	0.00	0.00591	0.00	0.00567
B	100	No Calibration	0.00	0.02044	0.00	0.01692
		Regression estimator	-0.01	0.01473	0.00	0.01461
		EL estimator ( $t = 1$ )	0.01	0.01652	0.01	0.01516
		EL estimator ( $t = 10$ )	0.00	0.01490	0.01	0.01472
		ET estimator ( $t = 1$ )	0.00	0.01516	0.01	0.01483
		ET estimator ( $t = 10$ )	0.00	0.01470	0.00	0.01459
		IVET estimator ( $t = 1$ )	0.00	0.01497	0.00	0.01458
		IVET estimator ( $t = 10$ )	0.00	0.01472	0.00	0.01453
	200	No Calibration	0.00	0.00888	0.00	0.00823
		Regression estimator	-0.01	0.00705	0.00	0.00735
		EL estimator ( $t = 1$ )	0.01	0.00769	0.01	0.00764
		EL estimator ( $t = 10$ )	0.00	0.00715	0.01	0.00745
		ET estimator ( $t = 1$ )	0.00	0.00723	0.01	0.00749
		ET estimator ( $t = 10$ )	0.00	0.00706	0.01	0.00734
		IVET estimator ( $t = 1$ )	0.00	0.00704	0.00	0.00728
		IVET estimator ( $t = 10$ )	0.00	0.00699	0.00	0.00725

SRS, simple random sampling; PPS, probability proportional to size sampling; MSE, mean squared error; EL, empirical likelihood; ET, exponential tilting; IVET, instrumental-variable exponential tilting.

Table 3

Monte Carlo Relative Biases of the variance estimators, based on 10,000 Monte Carlo samples

Population	Sample size	Estimator	Linearization		Replication	
			SRS	PPS	SRS	PPS
A	100	ET ( $t = 1$ )	-7.02	-2.66	10.65	4.11
		ET ( $t = 10$ )	-4.91	-0.80	5.60	0.67
		IVET ( $t = 1$ )	-5.28	-3.63	7.67	2.25
		IVET ( $t = 10$ )	-4.11	-0.87	4.96	0.41
	200	ET ( $t = 1$ )	-3.97	-0.19	3.65	0.57
		ET ( $t = 10$ )	-2.93	0.87	2.23	-0.35
		IVET ( $t = 1$ )	-3.35	-0.10	2.34	0.02
		IVET ( $t = 10$ )	-2.72	0.78	1.62	-0.53
B	100	ET ( $t = 1$ )	-7.64	-3.01	10.72	4.50
		ET ( $t = 10$ )	-5.98	-0.98	7.21	0.74
		IVET ( $t = 1$ )	-5.77	-2.31	4.53	-0.10
		IVET ( $t = 10$ )	-5.44	-1.86	5.17	-0.51
	200	ET ( $t = 1$ )	-2.41	-1.01	5.76	2.53
		ET ( $t = 10$ )	-1.29	0.18	4.30	1.91
		IVET ( $t = 1$ )	-1.39	-0.35	2.09	1.04
		IVET ( $t = 10$ )	-1.15	-0.06	2.04	0.99

SRS, simple random sampling; PPS, probability proportional to size sampling; ET, exponential tilting; IVET, instrumental-variable exponential tilting.

If a solution to (37) exists, it can be expressed as the limit of the form

$$w_{i(t)} \propto \prod_{s=0}^{t-1} \exp \{ -\hat{U}'_{(s)} \hat{\Sigma}_{aa(s)}^{-1} U(\mathbf{x}_i) \} \quad (38)$$

where  $\hat{U}_{(s)} = \sum_{i \in A} w_{i(s)} U(\mathbf{x}_i)$ ,  $\hat{\Sigma}_{aa(s)} = \sum_{i \in A} w_{i(s)} \{U(\mathbf{x}_i) - \bar{U}_{(0)}\}^{\otimes 2}$ ,  $\bar{U}_{(0)} = \sum_{i \in A} w_{i(0)} U(\mathbf{x}_i) / \sum_{i \in A} w_{i(0)}$  with the initial weight  $w_{i(0)} = d_i(\hat{N}/\hat{N}_d)$ . If the solution to condition (37) does not exist, we can still use the weights in (38), but the equality must be relaxed. Instead, approximate equality will be satisfied in (37) in the sense that  $\sum_{i \in A} w_{i(t)} U(\mathbf{x}_i)$  converges to zero much faster than  $\sum_{i \in A} w_{i(0)} U(\mathbf{x}_i)$  for  $t \geq 1$ . Approximate equality in (37) is called the approximate calibration condition.

The estimators  $\hat{Y}_{(t)} = \sum_{i \in A} w_{i(t)} y_i$  that use the  $t$ -step ET weights in (38), including the one-step estimator  $\hat{Y}_{(1)}$ , are asymptotically equivalent to the regression estimator of the form

$$\hat{Y}_{\text{reg}} = \hat{Y}_{(0)} - \hat{U}'_{(0)} \hat{\Sigma}_{aa(0)}^{-1} \hat{\Sigma}_{ay(0)},$$

where  $\hat{Y}_{(0)} = \sum_{i \in A} w_{i(0)} y_i$  and  $\hat{\Sigma}_{ay(0)} = \sum_{i \in A} w_{i(0)} \{U(\mathbf{x}_i) - \bar{U}_{(0)}\} y_i$ . Unlike the regression estimator, the weights of the proposed method are always nonnegative. Furthermore, using the instrumental variable technique in Section 3, the weights are bounded above. Suitable choice of the instrumental variable also improves the efficiency of the resulting calibration estimator.

The exponential tilting calibration method is asymptotically equivalent to the empirical likelihood calibration method but it is more attractive computationally in the sense that the partial derivatives are not required in the iterative computation. Because the computation is simple, the variance of the proposed estimator can be easily estimated using a replication method, as discussed in Section 4. Further investigation in this direction, including interval estimation, can be a topic of future research.

## Acknowledgements

The author wishes to thank Minsun Kim for computational support and two anonymous referees and the associated editor for very helpful comments that greatly improved the quality of the paper. This research was partially supported by a Cooperative Agreement NRCS 68-3A75-4-122 between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the USDA Natural Resources Conservation Service.

## Appendix

### A. Assumptions and proof of Theorem 1

We first assume the following regularity conditions:

[A-1] The density  $f(\mathbf{x}; \boldsymbol{\eta})$  is twice differentiable with respect to  $\boldsymbol{\eta}$  for every  $\mathbf{x}$  and satisfy

$$\left| \frac{\partial^2 f(\mathbf{x}; \boldsymbol{\eta})}{\partial \eta_i \partial \eta_j} \right| \leq K(\mathbf{x})$$

for function  $K(\mathbf{x})$  such that  $E\{K(\mathbf{x})\} < \infty$ , in a neighborhood of  $\boldsymbol{\eta}_{0,N}$ .

[A-2] The pseudo maximum likelihood estimator  $\hat{\boldsymbol{\eta}}$  satisfies  $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}) = O_p(1)$ .

[A-3] The matrix  $E\{\mathbf{s}(\boldsymbol{\eta}_{0,N})\}^{\otimes 2}$  exists and is nonsingular, where  $\mathbf{s}(\boldsymbol{\eta}_{0,N}) = \partial \ln f(\mathbf{x}_i; \boldsymbol{\eta}) / \partial \boldsymbol{\eta} |_{\boldsymbol{\eta}=\boldsymbol{\eta}_{0,N}}$ .

To prove Theorem 1, write

$$g_i(\boldsymbol{\eta}) = \frac{f(\mathbf{x}_i; \boldsymbol{\eta}_{0,N})}{f(\mathbf{x}_i; \boldsymbol{\eta})},$$

and  $w_i(\boldsymbol{\eta}) = d_i g_i(\boldsymbol{\eta})$ . The estimated importance weight in (8) can be written  $w_i = w_i(\hat{\boldsymbol{\eta}})$ . Taking a Taylor expansion of  $N^{-1} \sum_{i \in A} d_i \mathbf{s}_i(\hat{\boldsymbol{\eta}}) = 0$  around  $\boldsymbol{\eta}_{0,N}$  leads to

$$\begin{aligned} \mathbf{0} &= \frac{1}{N} \sum_{i \in A} d_i \mathbf{s}_i(\boldsymbol{\eta}_{0,N}) \\ &+ \left\{ \frac{\partial}{\partial \boldsymbol{\eta}'} \frac{1}{N} \sum_{i \in A} d_i \mathbf{s}_i(\boldsymbol{\eta}_{0,N}) \right\} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}) \\ &+ o_p(|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}|). \end{aligned}$$

Note that the first term on the right side of

$$\begin{aligned} \frac{1}{N} \frac{\partial}{\partial \boldsymbol{\eta}'} \sum_{i \in A} d_i \mathbf{s}_i(\boldsymbol{\eta}) &= \frac{1}{N} \sum_{i \in A} d_i \frac{\partial^2 f(\mathbf{x}_i; \boldsymbol{\eta}) / \partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'}{f(\mathbf{x}_i; \boldsymbol{\eta})} \\ &- \frac{1}{N} \sum_{i \in A} d_i \left\{ \frac{\partial f(\mathbf{x}_i; \boldsymbol{\eta}) / \partial \boldsymbol{\eta}}{f(\mathbf{x}_i; \boldsymbol{\eta})} \right\}^{\otimes 2}. \end{aligned} \quad (\text{A1})$$

converges to  $\int \{\partial^2 f(\mathbf{x}; \boldsymbol{\eta}) / \partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'\} d\mathbf{x}$  which equals to zero by the dominated convergence theorem with [A1]. The second term converges to  $E\{\mathbf{s}(\boldsymbol{\eta}_{0,N})\}^{\otimes 2}$ . Thus, by [A-2],

$$\bar{\mathbf{S}}_{0d} \equiv \frac{1}{N} \sum_{i \in A} d_i \mathbf{s}_i(\boldsymbol{\eta}_{0,N}) = O_p(n^{-1/2}) \quad (\text{A2})$$

and

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N} = \hat{\Sigma}_{ss}^{-1} \bar{\mathbf{S}}_{0d} + o_p(n^{-1/2}). \quad (\text{A3})$$



Now, taking a Taylor expansion of  $N^{-1}\hat{Y}_w = N^{-1}\sum_{i \in A} w_i(\hat{\eta}) y_i$  around  $\eta = \eta_{0,N}$  leads to

$$\frac{\hat{Y}_w}{N} = \frac{\hat{Y}_d}{N} + \left\{ \frac{\partial}{\partial \eta} \frac{1}{N} \sum_{i \in A} w_i(\eta_{0,N}) y_i \right\}' (\hat{\eta} - \eta_{0,N}) + o_p(|\hat{\eta} - \eta_{0,N}|) \quad (A4)$$

by the uniform continuity of  $\partial\{\sum_{i \in A} w_i(\eta) y_i\}/\partial\eta$  around  $\eta_{0,N}$ . Now, using

$$\frac{\partial}{\partial \eta} g_i(\eta) = \frac{f(\mathbf{x}_i; \eta)}{f(\mathbf{x}_i; \eta)} \cdot \frac{\partial f(\mathbf{x}_i; \eta)}{\partial \eta} = -g_i(\eta) \times s_i(\eta),$$

where  $s_i(\eta) = \partial \ln f(\mathbf{x}_i; \eta) / \partial \eta$ , we have

$$\frac{\partial}{\partial \eta} \sum_{i \in A} w_i(\eta) y_i = - \sum_{i \in A} w_i(\eta) s_i(\eta) y_i.$$

Using  $w_i(\eta_{0,N}) = d_i$  and writing  $s_i(\eta_{0,N}) = s_{i0}$ , we have, by (A2),

$$\begin{aligned} \frac{\partial}{\partial \eta} \frac{1}{N} \sum_{i \in A} w_i(\eta_{0,N}) y_i &= - \frac{1}{N} \sum_{i \in A} d_i s_{i0} y_i \\ &= - \hat{\Sigma}_{sy} + O_p(n^{-1/2}). \end{aligned} \quad (A5)$$

Using (A5) and (A3) in (A4), result (9) is obtained.

## B. Proof of Theorem 2

Write

$$\hat{\theta}(\lambda_1) = \frac{\sum_{i \in A} d_i m_i(\lambda_1) y_i}{\sum_{i \in A} d_i m_i(\lambda_1)},$$

where  $m_i(\lambda_1) = \exp(\lambda_1' \mathbf{x}_i)$ . Note that  $\hat{Y}_{ET(t)} = \hat{N} \hat{\theta}(\hat{\lambda}_{1(t)})$  and  $\hat{\lambda}_{1(t)}$  is defined in (19). By a Taylor expansion of  $\hat{\theta}(\hat{\lambda}_{1(t)}) = \hat{N}^{-1} \hat{Y}_{ET(t)}$  around  $\lambda_1 = \mathbf{0}$  and by the continuity of the partial derivatives of  $\hat{\theta}(\lambda_1)$ , we have

$$\hat{\theta}(\hat{\lambda}_{1(t)}) = \hat{\theta}(\mathbf{0}) + \hat{\theta}'(\mathbf{0})' (\hat{\lambda}_{1(t)} - \mathbf{0}) + o_p(|\hat{\lambda}_{1(t)} - \mathbf{0}|), \quad (B1)$$

where  $\hat{\theta}'(\lambda) = \partial \hat{\theta}(\lambda) / \partial \lambda$ . Because  $\hat{\lambda}_{1(t)}$  converges in quadratic order and the one-step estimator satisfies  $\hat{\lambda}_{1(t)} = O_p(n^{-1/2})$ , equation (22) can be written as

$$\begin{aligned} \hat{\lambda}_{1(t)} &= \left\{ \hat{N}^{-1} \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_d)' \right\}^{-1} (\hat{N} \hat{\mathbf{X}} - \bar{\mathbf{X}}_d) \\ &\quad + o_p(n^{-1/2}). \end{aligned} \quad (B2)$$

Note that

$$\hat{\theta}(\lambda_1) = \left\{ \sum_{i \in A} d_i m_i(\lambda_1) \right\}^{-1} \sum_{i \in A} d_i m_i(\lambda_1) \{y_i - \hat{\theta}(\lambda_1)\}$$

where  $\dot{m}_i(\lambda_1) = \partial m_i(\lambda_1) / \partial \lambda_1$ . Using  $m_i(\mathbf{0}) = 1$  and  $\dot{m}_i(\mathbf{0}) = \mathbf{x}_i$ , we have  $\hat{\theta}(\mathbf{0}) = \hat{Y}_d / \hat{N}_d$  and

$$\dot{\theta}(\mathbf{0}) = \hat{N}_d^{-1} \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_d) y_i. \quad (B3)$$

Therefore, inserting (B2) and (B3) into (B1), we have

$$\begin{aligned} \hat{\theta}(\hat{\lambda}_{1(t)}) &= \frac{\hat{Y}_d}{\hat{N}_d} \\ &\quad + \left( \frac{\mathbf{X}}{\hat{N}} - \bar{\mathbf{X}}_d \right)' \left\{ \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_d)^{\otimes 2} \right\}^{-1} \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_d) y_i \\ &\quad + o_p(n^{-1/2}), \end{aligned}$$

which proves (23).

## References

- Beaumont, J.-F., and Bocci, C. (2008). Another look at ridge calibration. *Metron*, LXVI, 5-20.
- Breidt, F.J., Claeskens, G. and Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92, 831-846.
- Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.
- Chen, J., and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.
- Chen, J., and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.
- Chen, J., Variyath, A.M. and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, 17, 426-443.
- Déville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Déville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedure in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9, 139-172.
- Estevao, V.M., and Särndal, C.-E. (2000). A functional approach to calibration. *Journal of Official Statistics*, 16, 379-399.
- Folsom, R.E. (1991). Exponential and logistic weight adjustment for sampling and nonresponse error reduction. In *Proceedings of the Section on Social Statistics*, American Statistical Association, 197-202.
- Fuller, W.A. (2002). Regression estimation for sample surveys. *Survey Methodology*, 28, 5-23.

- Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Givens, G.H., and Hoeting, J.A. (2005). *Computational Statistics*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Henmi, M., Yoshida, R. and Eguchi, S. (2007). Importance sampling via the estimated sampler. *Biometrika*, 94, 985-991.
- Imbens, G.W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20, 493-506.
- Isaki, C., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- Kim, J.K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, 19, 145-157.
- Kim, J.K., and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, In press.
- Kott, P.S. (2003). A practical use for instrumental-variable calibration. *Journal of Official Statistics*, 19, 265-272.
- Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.
- Kitamura, Y., and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65, 861-874.
- Park, M., and Fuller, W.A. (2009). The mixed model for survey regression estimation. *Journal of Statistical Planning and Inference*, 139, 1320-1331.
- Rao, J.N.K., and Singh, A. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. In *Proceedings of the Section on Survey Research Methods*, American Statist Association, 57-64.
- Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99-119.
- Särndal, C.-E., Swenson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer.
- Tillé, Y. (1998). Estimation in surveys using conditional probabilities: Simple random sampling. *International Statistical Review*, 66, 303-322.
- Wolter, K.M. (2007). *Introduction to Variance Estimation*. 2<sup>nd</sup> Ed. New York: Springer-Verlag.
- Wu, C., and Rao, J.N.K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*, 34, 359-375.





# Comparison of survey regression techniques in the context of small area estimation of poverty

Stephen J. Haslett, Marissa C. Isidro and Geoffrey Jones<sup>1</sup>

## Abstract

One key to poverty alleviation or eradication in the third world is reliable information on the poor and their location, so that interventions and assistance can be effectively targeted to the neediest people. Small area estimation is one statistical technique that is used to monitor poverty and to decide on aid allocation in pursuit of the Millennium Development Goals. Elbers, Lanjouw and Lanjouw (ELL) (2003) proposed a small area estimation methodology for income-based or expenditure-based poverty measures, which is implemented by the World Bank in its poverty mapping projects via the involvement of the central statistical agencies in many third world countries, including Cambodia, Lao PDR, the Philippines, Thailand and Vietnam, and is incorporated into the World Bank software program PovMap. In this paper, the ELL methodology which consists of first modeling survey data and then applying that model to census information is presented and discussed with strong emphasis on the first phase, *i.e.*, the fitting of regression models and on the estimated standard errors at the second phase. Other regression model fitting procedures such as the General Survey Regression (GSR) (as described in Lohr (1999) Chapter 11) and those used in existing small area estimation techniques: Pseudo-Empirical Best Linear Unbiased Prediction (Pseudo-EBLUP) approach (You and Rao 2002) and Iterative Weighted Estimating Equation (IWEE) method (You, Rao and Kovačević 2003) are presented and compared with the ELL modeling strategy. The most significant difference between the ELL method and the other techniques is in the theoretical underpinning of the ELL model fitting procedure. An example based on the Philippines Family Income and Expenditure Survey is presented to show the differences in both the parameter estimates and their corresponding standard errors, and in the variance components generated from the different methods and the discussion is extended to the effect of these on the estimated accuracy of the final small area estimates themselves. The need for sound estimation of variance components, as well as regression estimates and estimates of their standard errors for small area estimation of poverty is emphasized.

Key Words: Small area models; Nested error regression model; Poverty mapping.

## 1. Introduction

Poverty is a very complex multidimensional concern: there is no single definition and method of measurement available. In this paper, we adhere to the meaning of poverty that is used by most economists, *i.e.*, households are considered to be in poverty if their income falls below some income threshold called the poverty line. Chambers (2006) described this as income-poverty, and it is the definition adopted by the World Bank in the implementation of their small area poverty mapping projects carried out in conjunction with national statistical agencies and used, for example, for monitoring progress towards the Millennium Development Goals (UN website). Sometimes expenditure-based poverty measures are used instead to assess economic poverty. In public health related contexts, different measures such as standardized weight for age, height for age and weight for height for children (underweight, stunting and wasting, respectively) are used, *e.g.*, in Bangladesh (Haslett and Jones 2004) and Nepal (Haslett and Jones 2006).

Surveys conducted in most third world countries usually allow an acceptable level of precision for reporting poverty statistics at the first and second administrative level or geographical area (*e.g.*, for the Philippines - National and

Region respectively). However, for policy makers to properly target assistance and interventions to the neediest communities and households, more disaggregated finer-level poverty statistics are needed. However, survey based poverty statistics at smaller geographical areas or lower administrative level are usually less reliable (have higher standard errors) due to smaller sample sizes, and this is where small area estimation comes into play.

The most common small area estimation methodology used for poverty measures in third world countries proposed by Elbers, Lanjouw and Lanjouw (ELL) (2002, 2003) allows generation of more precise estimates for smaller geographical areas by combining the survey data with information from a recent census. The ELL method consists of two phases: fitting a regression model (or models) to complex survey data and using that model to predict income or expenditure per capita at household level (which is transformed and aggregated to estimate poverty statistics at small area level).

In this paper, we focus specifically on the various algorithms used to fit the phase 1 regression models, and to estimate regression parameter standard errors and variance components from survey data. We emphasise consequences of survey regression modeling decisions rather than the

1. Stephen J. Haslett, Marissa C. Isidro and Geoffrey Jones, Institute of Fundamental Sciences: Statistics, College of Sciences, Massey University, Private Bag 11-222, Palmerston North, New Zealand. E-mail: S.J.Haslett@massey.ac.nz.

entire and rather comprehensive system ELL use to form small area estimates.

The preliminary requirement of the ELL methodology applied to economic measures is to develop an accurate model of per capita income or expenditure of households although this is often used to generate non-linear functions of income or expenditure (*e.g.*, poverty incidence - percentage of households below the poverty line, or poverty gap - sum of relative differences in income or expenditure for households or individuals below the poverty line). The survey-based regression model developed for income or expenditure is critical to accurate poverty statistics, but as we show below the regression model itself is not always the most important element, and other issues such as estimation of variance components deserve emphasis.

Other existing survey-based small area estimation regression techniques - Pseudo-Empirical Best Linear Unbiased Prediction (Pseudo-EBLUP) approach (You and Rao 2002), Iterative Weighted Estimating Equation (IWEE) method (You *et al.* 2003) and the General Survey Regression (GSR) (Skinner, Holt and Smith 1989) method are considered as alternative survey based model-fitting techniques and compared with two variations of the ELL method for fitting regression models to survey data. Our investigation is based on real data from the 2000 Philippine Family Income and Expenditure Survey (FIES), rather than simulated data.

This paper is organized as follows: Section 2 gives relevant background on small area models; the model for income (or expenditure) as presented by Elbers, Lanjouw and Lanjouw is given in Section 3; presented in Section 4 is a summary of the ELL methodology, followed by details on the alternative fitting methods in Section 5, which includes the Pseudo-Empirical Best Linear Unbiased Prediction Approach (5.1), IWEE Method (5.2), and the General Survey Regression Method (5.3). Section 6 discusses differences between the techniques, while Section 7 presents their application to the Philippine FIES 2000 data. This is followed by the conclusion and recommendations (Section 8).

## 2. Small area models

Ghosh and Rao (1994) classify small area models into two broad categories, area level and unit level models. Area level models refer to sets of models that can be considered when only area-specific auxiliary variables are available. Unit level models, on the other hand, refer to models that can be considered when there are unit-specific auxiliary variables and unit level values of the variable under study can be used. All such models are special cases of a general linear or generalized linear mixed model, and usually involve both fixed and random effects.

For area level models, it is assumed that the population mean ( $\bar{Y}_a$ ) of the  $a^{\text{th}}$  small area or some suitable function  $\theta_a = g(\bar{Y}_a)$  is related to the area-specific auxiliary variables  $\mathbf{x}_a = (x_{a1}, \dots, x_{ap})'$  through a linear model

$$\theta_a = \mathbf{x}_a' \boldsymbol{\beta} + c_a v_a \quad (1)$$

where  $a = 1, \dots, k$ ,  $v_a \sim \text{iid}(0, \sigma_v^2)$ ,  $\boldsymbol{\beta}$  is a vector of regression parameters,  $c_a$  are known or estimated positive constants to allow for heteroscedasticity,  $k$  is the total number of small areas under study and  $p$  is the number of auxiliary variables. It is assumed that a direct design-based estimator,  $\hat{\bar{Y}}_a$ , of the population mean  $\bar{Y}_a$  is available whenever the area sample size  $n_a \geq 1$ , and that

$$\hat{\theta}_a = \theta_a + e_a \quad (2)$$

where  $\hat{\theta}_a = g(\hat{\bar{Y}}_a)$  and the sampling errors  $e_a$  are independent  $N(0, V_a)$  with known variance  $V_a$ . Combining equation (1) and (2) gives the area level linear mixed model:

$$\hat{\theta}_a = \mathbf{x}_a' \boldsymbol{\beta} + c_a v_a + e_a. \quad (3)$$

We note that (3) involves both design-based random variables  $e_a$  and model-based random variables  $v_a$  (Rao 1999), where design-based variables are due to the sample selection mechanism, and model-based ones to the super-population structure in which the model is embedded.

Area level models have various extensions so they can for example handle correlated sampling errors, spatial dependence of random small area effects, time series and cross-sectional data (see Rao 2003, 1999 and Ghosh and Rao 1994).

The unit level model assumes that the variable of interest  $Y_{ah}$  for the  $h^{\text{th}}$  unit in the  $a^{\text{th}}$  small area is related to the element-specific auxiliary data  $\mathbf{x}_{ah} = (x_{ah1}, \dots, x_{ahp})'$  through a nested error regression model:

$$Y_{ah} = \mathbf{x}_{ah}' \boldsymbol{\beta} + v_a + e_{ah} \quad (4)$$

where  $a = 1, \dots, k$ ,  $h = 1, \dots, N_a$ ,  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p-1})$  is  $p \times 1$  vector of regression parameters and  $N_a$  is the number of population units or households in the  $a^{\text{th}}$  small area. It is also assumed that the random effects  $v_a$  are  $\text{iid} N(0, \sigma_v^2)$  and are independent of the unit errors  $e_{ah}$  which are assumed to be  $\text{iid} N(0, \sigma_e^2)$ . Extensions that allow errors to be heteroscedastic, with known scaling constant(s) are also possible.

The ELL method uses a unit level model, where the units are households in the case of income or expenditure data, and where the variation is modeled at primary sampling unit, *i.e.*, cluster level and household level. Note that ELL do not include model variation at small area level, only for cluster within small area, and for household within cluster. This is the form of the basic model used for comparisons in this paper since ELL is the standard small area estimation



method for poverty in third world countries. In the real datasets we have studied this additional small area variation has been very small. Despite this empirical evidence however, important questions remain about how best to estimate the small area variance component in the presence of cluster level variation, when there is sample survey weighting, especially where many of the small areas contain only one sampled cluster.

The ELL model has a number of other characteristics not all of which are standard in a statistical sense (see Haslett and Jones 2005, for example). The intention of this paper is not to discuss differences in the available methods generally, but to focus directly on how methods of fitting regression models to survey data differ when the ELL first phase “base structure” of fitting a survey regression model is used. The focus of this paper therefore is on comparison of the available methods of fitting regression models to survey data on income or expenditure using a specified set of regressors, even though ELL can also be (and is) used relatively routinely to find small area estimates for non-linear functions (e.g., poverty incidence, gap or severity) by applying fitted regression models to a census.

The answer to the ‘best regression model fitting’ question for survey data on which this paper focuses (as with other matters related to the ELL methodology) is particularly important because there are billions of dollars of aid funding that are (or have the potential to be) allocated based on the regression models used as part of small area estimation of poverty.

### 3. Income/consumption model

Modeling per capita income or expenditure of households instead of poverty measures themselves (such as poverty incidence and gap) is one of the distinctive features of the ELL method. As mentioned in the previous section, the ELL method involves fitting the income or expenditure model to the survey data and applying it to the census data prior to the generation of the small area estimates of poverty measures. The income/expenditure model is as follows:

$$Y_{bh} = \mathbf{x}'_{bh}\boldsymbol{\beta} + u_{bh} \quad (5)$$

where  $b = 1, \dots, M$ ,  $h = 1, \dots, N_b$ ;  $Y_{bh}$  is the log-transformed per capita income or expenditure of the  $h^{\text{th}}$  unit or household in the  $b^{\text{th}}$  cluster,  $M$  is the total number of clusters in the population and  $N_b$  is the total number of households in the  $b^{\text{th}}$  cluster in the population.  $\mathbf{x}_{bh}$  is a set of the auxiliary variables available in both the survey and the census, which generally need to be contemporaneous;  $u_{bh}$  is the random error term representing that part of  $Y_{bh}$  that cannot be explained by  $\mathbf{x}_{bh}$ . Income and expenditure

data almost invariably have a skewed distribution, hence a transformation (usually logarithmic) is applied to make the data more symmetrical.

The households for which data on per capita income or expenditure is collected are seldom independent, but have natural groupings or clusters, often defined administratively. Households that are close to each other or in the same cluster, tend to be similar in many respects. In the survey data, the clusters are usually also the primary sampling units (PSUs) for the sample survey design. To account for the clustering of households, the random error term  $u_{bh}$  in the regression model is usually assumed to have the following specification:

$$u_{bh} = v_b + e_{bh} \quad (6)$$

where  $v$  and  $e$  are independent of each other and uncorrelated with  $\mathbf{x}_{bh}$ ,  $v_b$  is the error term held in common by the  $b^{\text{th}}$  group or cluster (e.g., barangay for the Philippines) and  $e_{bh}$  is the household level error within the cluster. The importance of each term is measured by their respective variances or variance components,  $\sigma_v^2$  and  $\sigma_e^2$ . There are various procedures for estimating these variances. This important topic is covered in the sections that follow.

Model (5) can be written as

$$Y_{bh} = \mathbf{x}'_{bh}\boldsymbol{\beta} + v_b + e_{bh} \quad (7)$$

which is similar in form to the unit level model or nested error regression model mentioned in the previous section. However while the form of the model is similar, the group being referred to is different, e.g.,  $Y_{ah}$  refers to the  $h^{\text{th}}$  household in the  $a^{\text{th}}$  small area, while  $Y_{bh}$  refers to the  $h^{\text{th}}$  household in the  $b^{\text{th}}$  cluster. Clusters, based on the survey design, will typically be much smaller than the areas for which small area estimates are sought, and generally (unlike almost all the small areas) not all clusters are sampled. For example in the Philippines, estimates are sought at the municipal level which is composed of barangays or clusters.

### 4. The ELL methodology

In the ELL methodology, the estimate of the regression parameter  $\boldsymbol{\beta}$  is given, in Elbers *et al.* (2002, page 11 footnote 8) and in the POVMAP software Zhao (2006) developed for the ELL method, as

$$\hat{\boldsymbol{\beta}}_{\text{ELL}} = \left( \sum_{b=1}^m \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{X}_b \right)^{-1} \left( \sum_{b=1}^m \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{y}_b \right) \quad (8)$$

and the corresponding variance-covariance matrix as

$$\mathbf{V}(\hat{\boldsymbol{\beta}}_{\text{ELL}}) = \mathbf{D} \left[ \left( \sum_{b=1}^m \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{W}_b \mathbf{X}_b \right)^{-1} \right] \mathbf{D} \quad (9)$$



where  $\mathbf{V}_b = (\sigma_e^2 \mathbf{I}_{n_b} + \sigma_v^2 \mathbf{1}_{n_b} \mathbf{1}'_{n_b})$ ,  $(\sigma_v^2)$  is the cluster level variance, while  $(\sigma_e^2)$  is the household level variance,  $\mathbf{I}_{n_b}$  is an identity matrix,  $\mathbf{1}'_{n_b} = (1 \dots 1)$  is a constant vector,  $\mathbf{D} = (\sum_{b=1}^m \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{X}_b)^{-1}$ ,  $\mathbf{X}_b = (\mathbf{x}_{b1}, \dots, \mathbf{x}_{bn_b})'$ ,  $\mathbf{y}_b = (y_{b1}, \dots, y_{bn_b})'$ ;  $\mathbf{W}_b$  is a diagonal matrix of sampling weights;  $m$  is the number of clusters in the sample and  $n_b$  is the number of households in each sampled cluster. Equation (8) assumes  $\mathbf{V}_b$  is known. In practice we need to estimate  $\sigma_e^2$  and  $\sigma_v^2$  to get the estimator  $\hat{\mathbf{V}}_b$ . We note that the variance expression in (9) is derived under a vaguely specified model assumed for the sample (see Elbers *et al.* 2002). Under the ELL method, fitting the income/expenditure model (7) involves obtaining the initial estimate of  $\beta$  through weighted least squares (WLS) method and using the residuals of the initial model to estimate the covariance matrix  $\mathbf{V}_b$  needed to obtain  $\hat{\beta}_{\text{ELL}}$ . The estimate of the cluster level  $(\sigma_v^2)$  and household level  $(\sigma_e^2)$  variances, are derived by Elbers *et al.* (2002) as follows:

$$\hat{\sigma}_v^2 = \max \left( \frac{\sum_b w_b (u_b - u_{..})^2}{\sum_b w_b (1 - w_b)} - \frac{\sum_b w_b (1 - w_b) \tau_b^2}{\sum_b w_b (1 - w_b)}; 0 \right) \quad (10)$$

where  $\tau_b^2 = \sum_h (e_{bh} - e_b)^2 / (n_b (n_b - 1))$ ;  $w_b = \sum_h w_{bh} / \sum_b \sum_h w_{bh}$ , is the by-cluster transformed sampling weights which sum to one across clusters and  $w_{bh}$  is the re-scaled sampling weights which sum to the total sample size. Here  $u_b = \sum_h u_{bh}$  and  $u_{..} = \sum_b \sum_h u_{bh}$  (which is equal to zero) where  $u_{bh}$  is as defined in equation (6).

There are two ways suggested by Elbers *et al.* (2002) to generate the estimate of the household level variance component: "direct" computation which is denoted by  $(\hat{\sigma}_e^2)$  or the heteroscedasticity model-based  $(\hat{\sigma}_{e,bh}^2)$ . Direct computation involves using the difference between the estimated mean square error from the initial WLS regression and the computed estimate of  $\sigma_v^2$ , while the heteroscedasticity model-based computation uses a logistic-type link function to bound the variance as follows:

$$\sigma_{e,bh}^2(\mathbf{z}_{bh}, \alpha, A, B) = \left[ \frac{A \exp(\mathbf{z}'_{bh} \alpha) + B}{1 + \exp(\mathbf{z}'_{bh} \alpha)} \right] \quad (11)$$

where A and B are the upper and lower bounds respectively, estimated with the parameter vector  $\alpha$  using a standard pseudomaximum likelihood procedure (Elbers *et al.* 2003), and where  $\mathbf{z}_{bh}$  are auxiliary variables. Elbers *et al.* claim that imposing a minimum bound of zero and a maximum bound of  $A^* = (1.05) \max\{e_{bh}^2\}$  in general yields similar estimates of the parameters  $\alpha$ . These restrictions allow one to estimate the simpler form

$$\ln \left[ \frac{e_{bh}^2}{A^* - e_{bh}^2} \right] = \mathbf{z}'_{bh} \alpha + r_{bh} \quad (12)$$

where  $r_{bh}$  is an error term and the other variables are as defined earlier. In most of the World Bank poverty mapping projects, slight modifications are usually made, for example, adding a constant  $\delta$  to  $e_{bh}^2$  in model (11).

By using model (12), and employing the delta method,  $\hat{\sigma}_{e,bh}^2$  is computed as:

$$\hat{\sigma}_{e,bh}^2 = \left[ \frac{A^* C_{bh}}{1 + C_{bh}} \right] + \frac{1}{2} \hat{\sigma}_r^2 \left[ \frac{A^* C_{bh} (1 - C_{bh})}{(1 + C_{bh})^3} \right] \quad (13)$$

where  $C_{bh} = \exp\{\mathbf{z}'_{bh} \hat{\alpha}\}$ , and  $\hat{\sigma}_r^2$  is the estimated variance of the residuals under model (12). If the household level variance component is based on a heteroscedastic model, then,  $\mathbf{V}_b = (\sigma_{e,bh}^2 \mathbf{I}_{n_b} + \sigma_v^2 \mathbf{1}_{n_b} \mathbf{1}'_{n_b})$ . Heteroscedasticity modeling is conducted on the assumption that variation at the household level depends on some covariates.

As discussed in more detail in the appendix, the way in which the weight matrix  $\mathbf{W}_b$  enters the calculation in equation (9) above leads to an asymmetric estimated covariance matrix. A rather better approach based on 'pseudomaximum likelihood' is outlined by Pfeiffermann, Skinner, Holmes, Goldstein and Rasbash (1998) and involves splitting  $\mathbf{X}'_b \mathbf{V}_b^{-1} \mathbf{X}_b$  into separate sums of squares and cross-product terms, and weighting each appropriately - if we write  $\mathbf{V}_b^{-1} = c \mathbf{I}_{n_b} + d \mathbf{1}_{n_b} \mathbf{1}'_{n_b}$  then the appropriate weighting is  $c \mathbf{X}'_b \mathbf{W}_b \mathbf{X}_b + d \mathbf{X}'_b \mathbf{W}_b \mathbf{1}_{n_b} \mathbf{1}'_{n_b} \mathbf{W}_b \mathbf{X}_b$ .

Since the ELL version,  $\mathbf{W}_b \mathbf{V}_b^{-1}$ , is not generally symmetric, neither is  $\mathbf{D}$  in equation (9). As a consequence the supposed covariance matrix of  $\hat{\beta}_{\text{ELL}}$ ,  $\mathbf{V}(\hat{\beta}_{\text{ELL}})$ , is also not symmetric. The POVMAP software attempts to solve this problem by taking the average of their  $\mathbf{V}(\hat{\beta}_{\text{ELL}})$  and its transpose, thereby forcing the matrix to be symmetric.

Note again that under the ELL method, the regression fit to the survey data and the estimation of variance components is only the first phase. The consequent phase involves prediction at household level based on the entire census data and aggregation to small area level.

The survey fitting methods (derivation of the estimate of  $\beta$  and its corresponding variance-covariance matrix) of three alternative regression procedures to ELL are presented in the following sections.

## 5. Alternative fitting methods

### 5.1 The pseudo-empirical best linear unbiased prediction approach

You and Rao (2002) proposed an estimator of the small area mean by deriving an estimator of  $\beta$  based on the unit level model (4). The process of deriving the estimator of  $\beta$  starts with the computation of the best linear unbiased predictor (BLUP) of  $v_a$  given the parameters  $\beta$ ,  $\sigma_e^2$  and

$\sigma_v^2$  from the aggregated (survey-weighted) area level model:

$$\bar{Y}_{aw} = \bar{\mathbf{x}}'_{aw} \boldsymbol{\beta} + v_a + \bar{e}_{aw} \quad (14)$$

which proceeds as follows:

$$\hat{v}_{aw}(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2) = \gamma_{aw}(\bar{Y}_{aw} - \bar{\mathbf{x}}'_{aw} \boldsymbol{\beta}) \quad (15)$$

where  $\bar{\mathbf{x}}_{aw} = \sum_{h=1}^{n_a} w_{ah} \mathbf{x}_{ah}$ ,  $\bar{Y}_{aw} = \sum_{h=1}^{n_a} w_{ah} y_{ah}$ ,  $\gamma_{aw} = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2 \delta_a^2)$ ,  $w_{ah} = \tilde{w}_{ah} / \sum_{h=1}^{n_a} \tilde{w}_{ah}$ ,  $\delta_a^2 = \sum_{h=1}^{n_a} w_{ah}^2$ , and  $\tilde{w}_{ah}$  are the unit level survey weights; then solving for the survey-weighted estimating equation for  $\boldsymbol{\beta}$ :

$$\sum_{a=1}^k \sum_{h=1}^{n_a} \tilde{w}_{ah} \mathbf{x}_{ah} [y_{ah} - \mathbf{x}'_{ah} \boldsymbol{\beta} - \hat{v}_{aw}(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2)] = 0 \quad (16)$$

from which the estimator of  $\boldsymbol{\beta}$  is obtained as

$$\hat{\boldsymbol{\beta}}_w = \left\{ \sum_{a=1}^k \sum_{h=1}^{n_a} \mathbf{x}_{ah} \mathbf{z}'_{ah} \right\}^{-1} \left\{ \sum_{a=1}^k \sum_{h=1}^{n_a} \mathbf{z}_{ah} y_{ah} \right\} \quad (17)$$

where  $\mathbf{z}_{ah} = \tilde{w}_{ah}(\mathbf{x}_{ah} - \gamma_{aw} \bar{\mathbf{x}}_{aw})$ . The corresponding covariance matrix is then as follows:

$$\begin{aligned} \Phi_w = \sigma_e^2 & \left( \sum_{a=1}^k \sum_{h=1}^{n_a} \mathbf{x}_{ah} \mathbf{z}'_{ah} \right)^{-1} \\ & \left( \sum_{a=1}^k \sum_{h=1}^{n_a} \mathbf{z}_{ah} \mathbf{z}'_{ah} \right) \left( \sum_{a=1}^k \sum_{h=1}^{n_a} \mathbf{x}_{ah} \mathbf{z}'_{ah} \right)^{-1} \\ & + \sigma_v^2 \left( \sum_{a=1}^k \sum_{h=1}^{n_a} \mathbf{x}_{ah} \mathbf{z}'_{ah} \right)^{-1} \\ & \left\{ \sum_{a=1}^k \left( \sum_{h=1}^{n_a} \mathbf{z}_{ah} \right) \left( \sum_{h=1}^{n_a} \mathbf{z}_{ah} \right)' \right\} \left\{ \left( \sum_{a=1}^k \sum_{h=1}^{n_a} \mathbf{x}_{ah} \mathbf{z}'_{ah} \right)^{-1} \right\}'. \quad (18) \end{aligned}$$

The variance components are estimated using Henderson's Method 3 (Henderson 1953), to generate unbiased estimates even in the presence of correlated elements in the model. The estimators of the variance components are as follows:

$$\hat{\sigma}_{eH}^2 = (n - k - p + 1)^{-1} \sum_{a=1}^k \sum_{h=1}^{n_a} \hat{e}_{ah}^2 \quad (19)$$

where  $\{\hat{e}_{ah}^2\}$  are residuals from the OLS regression of  $(y_{ah} - \bar{y}_a)$  on  $\{x_{ah1} - \bar{x}_{a,1}, \dots, x_{ahp} - \bar{x}_{a,p}\}$  and  $(\bar{y}_a, \bar{x}_{a,1}, \dots, \bar{x}_{a,p})$  are the sample means in the  $a^{\text{th}}$  group.

$$\hat{\sigma}_{vH}^2 = n_*^{-1} \left[ \sum_{a=1}^k \sum_{h=1}^{n_a} \hat{u}_{ah}^2 - (n - p) \hat{\sigma}_{eH}^2 \right] \quad (20)$$

where  $n_* = n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1} \sum_{a=1}^k n_a^2 \bar{\mathbf{x}}_a \bar{\mathbf{x}}_a']$  with  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ , and the  $\{\hat{u}_{ah}\}$  are the residuals from the OLS regression of  $y_{ah}$  on  $\{x_{ah1}, \dots, x_{ahp}\}$ . For the model (7), the subscript  $a$  is replaced by  $b$ .

However, the Henderson's estimators above do not account for the sampling weights. To address this, an estimation technique has been proposed by You *et al.* (2003) which extends the Pseudo-EBLUP method by incorporating the weights in the estimation of the variance components. This is described in the next section.

## 5.2 The iterative weighted estimating equation method

The estimator proposed by You *et al.* (2003) is similar to the Pseudo-EBLUP estimator, except that it incorporates the sampling weights in the computation of the variance components, and it generates the parameter estimate  $\boldsymbol{\beta}$  and the variance components by using an iterative weighted estimating equation (IWEE) approach. The authors derived the estimator of  $\sigma_e^2$  and  $\sigma_v^2$  as follows:

$$\begin{aligned} \hat{\sigma}_{ew}^{2(t)} &= \frac{\sum_{a=1}^k \sum_{h=1}^{n_a} \tilde{w}_{ah} [y_{ah} - \bar{Y}_{aw} - (\mathbf{x}_{ah} - \bar{\mathbf{x}}_{aw})' \hat{\boldsymbol{\beta}}^{(t-1)}]^2}{\sum_{a=1}^k \left[ (1 - \delta_a^2) \sum_{h=1}^{n_a} \tilde{w}_{ah} \right]} \\ &\equiv \tilde{\sigma}_{ew}^{2(t)}(\boldsymbol{\beta}) \quad (21) \end{aligned}$$

and

$$\begin{aligned} \hat{\sigma}_{vw}^{2(t)} &= \frac{1}{k} \sum_{a=1}^k \tilde{r}_{aw}^2 + \frac{\tilde{\sigma}_{vw}^{2(t-1)}}{k} \sum_{a=1}^k (\gamma_{aw} - 1)^2 + \frac{\tilde{\sigma}_{ew}^{2(t)}}{k} \sum_{a=1}^k \delta_a^2 \gamma_{aw}^2 \\ &\equiv \tilde{\sigma}_{vw}^{2(t)}(\tilde{v}_w, \sigma_e^2, \sigma_v^2). \quad (22) \end{aligned}$$

The survey weighted estimates of  $\boldsymbol{\beta}$ ,  $\sigma_e^2$ ,  $\sigma_v^2$  are obtained simultaneously by following iterative updating steps,  $t$  in the equation above stands for the  $t^{\text{th}}$  iteration. Since the variance components  $\sigma_v^2$  and  $\sigma_e^2$  are unknown, initial estimates for the iterative steps are generated by Henderson's method. Again, as for Pseudo-EBLUP, for the ELL regression model formulation (7), the subscript  $a$  is replaced by  $b$ .

This approach is similar to the probability-weighted iterative generalized least squares (PIWGLS) method proposed by Pfeiffermann *et al.* (1998) for fitting multilevel models where the estimation process considered the unequal selection probabilities at each stage of sampling and involves iterating between the parameter  $\boldsymbol{\beta}$  and the variance components until convergence. A model-based approach is also proposed by Pfeiffermann, Moura and Silva (2006), which involves deriving the hierarchical model for given sample data as a function of the population model and the selection probabilities, and then fitting the sample model using Bayesian approach by use of Markov Chain Monte Carlo algorithm.

## 5.3 General survey regression method

Another approach to generate the estimator of the parameter  $\boldsymbol{\beta}$  and its variance is the design-based methodology for fitting regression models (Lohr 1999). This



technique is currently used in the Stata, Sudaan, and WesVar package, for example. The estimator of  $\beta$  given below is the sample weighted regression estimator for a model with homoscedastic variance structure and uncorrelated observations in the population.

$$\hat{\beta}_s = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\mathbf{y}. \quad (23)$$

This estimator is not derived under the model specified by (7) even under the homoscedastic variances for household errors. The linearized/robust variance estimate for  $\hat{\beta}_s$  is based on the design-based variance estimator for a total, given as,

$$\hat{\mathbf{V}}(\hat{\beta}_s) = \mathbf{D} \left\{ \frac{m}{m-1} \sum_{b=1}^m \left( \sum_{h=1}^{n_b} w_{bh} \mathbf{d}_{bh} \right)' \left( \sum_{h=1}^{n_b} w_{bh} \mathbf{d}_{bh} \right) \right\} \mathbf{D} \quad (24)$$

where  $\mathbf{d}_{bh} = \hat{e}_{bh} \mathbf{x}_{bh}$ ;  $\hat{e}_{bh}$  is the residual from WLS regression;  $\mathbf{x}_{bh}$  is a vector of the independent variables;  $w_{bh}$  is a sampling weight;  $\mathbf{D} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ ; and  $\mathbf{W}$  is a diagonal matrix of the sampling weights.

The General Survey Regression method differs from the other techniques in the computation of the estimates, and generates the estimates without computing the variance components,  $\sigma_v^2$  and  $\sigma_e^2$ . As shown above, the equations for the estimator of the parameter  $\beta$  and its corresponding estimated covariance matrix only involve the sampling weights matrix  $\mathbf{W}$ . The estimated covariance matrix in (24) is often referred to as a sandwich estimator.

## 6. Comparison of the model fitting techniques

The ELL methodology is claimed to be a weighted GLS estimation procedure. However, as pointed out earlier, the sampling weights are not properly incorporated in the estimation process and this leads to non-interpretability of the elements in some matrices involved in the estimation, as well as asymmetry in the estimated covariance matrix. For the ELL method of estimating the variance components, the weights are accounted for only at the cluster level. The two ways (direct computation and heteroscedasticity model-based) that ELL use for generating the household level variance component do not incorporate the sampling weights. For direct computation, the household level variance component is determined from the residual of the survey-weighted (WLS) regression conducted at the preliminary step and the weighted estimate of the cluster level component. The heteroscedasticity based computation is based on modeling the square of the residuals from the WLS regression.

While the ELL methodology follows a GLS-like estimation procedure, the pseudo-EBLUP and IWEE method

follow the Generalized Estimating Equation (GEE) procedure (Liang and Zeger 1986) using an exchangeable working correlation matrix, *i.e.*, all the off-diagonal elements of the correlation matrix within clusters are equal, and in Pseudo-EBLUP and IWEE are equal to  $\sigma_v^2/(\sigma_v^2 + \sigma_e^2)$ . An exchangeable or equicorrelated working correlation matrix is one of the common working correlation matrices presented in the paper of Horton and Lipsitz (1999) when reviewing different software for fitting GEE regression models.

The two procedures, Pseudo-EBLUP and IWEE, both incorporate the sampling weights in the estimation of the parameter  $\beta$  and the corresponding standard error, although the Pseudo-EBLUP method uses Henderson's method in the estimation of the variance components. While Henderson's method generates unweighted estimates of the variance components, the IWEE method incorporates the sampling weights iteratively from estimation of variance components for computation of standard error of the estimate of the regression parameter.

There is a very limited published literature on the application to real data sets of the Pseudo-EBLUP and IWEE methods. Those that there are consider the clusters as the small area, and often use the data in Battese, Harter and Fuller (1988), whose data set contains information on hectares of corn and soybeans per segment for counties in North Central Iowa and assumes simple random sampling within areas or clusters. An exception is the recent paper by Militino, Ugarte, Goicoa and Gonzalez-Audicana (2006), which applies Pseudo-EBLUP to estimating the total area occupied by olive trees in Navarra, Spain, where (as in Battese *et al.*) the units are self weighting. Generally for poverty estimation, Pseudo-EBLUP and IWEE techniques must be applied in more complex situations, since sampling clusters and small areas are not identical and the sample is not self weighting. In the example in the next section, the clusters (barangay) are different from the small areas (municipalities), the clusters are sub-units of the small area and the sampling scheme is not self weighting.

The GSR method is one of the least complicated estimation procedures as it employs a weighted least squares procedure using the sandwich estimator for estimating the variance of the estimator of the regression parameter. As mentioned earlier, this method differs from the other techniques in that the estimate of the regression parameters and their corresponding standard errors are generated without computing the variance components.

Based on the discussion above, for all the techniques considered, the survey-based estimation procedure for the parameter  $\beta$  and its corresponding standard error are theoretically sound given their assumptions, except for the ELL method where there are some inconsistencies in the estimation of parameters  $\beta$  and the covariance of  $\hat{\beta}$ .



## 7. Application to real data

In this section, the four different regression techniques (one of which contains two variants of ELL) are compared using the Philippine 2000 Family Income and Expenditure Survey (FIES). The FIES data is a nationwide survey undertaken by the Philippines National Statistics Office (NSO) every three years. The survey gathers details on family income and expenditure as well as information affecting income and expenditure. Selected households are interviewed in two separate operations, each covering a half-year period, in order to allow for seasonal patterns in income and expenditure. For FIES 2000 the interviews were conducted in July 2000, for the period 01 January to 30 June and January 2001 for the period 01 July to 31 December. The sample design for FIES used a multi-stage stratified random sampling technique. Barangays are the primary sampling units (PSUs) and are stratified into urban and rural within each province and selected using systematic sampling with probability proportional to size. Large barangays are further divided into enumeration areas and subjected to further sampling before the final stage in which households are systematically sampled from the 1995 Population Census List of Households. Interview non-response was only 3.4 percent, with 39,615 of the sample households being successfully interviewed in both survey visits. Deterministic imputation was done to address item non-response, *i.e.*, entry for a particular missing item is deduced from other items in the questionnaire.

The auxiliary variables used in this paper are adopted from the variables included in the model formulated by Haslett and Jones (2005) that was fitted without using POVMAP for the small area poverty mapping project in the Philippines. The auxiliary variables included both household characteristics and municipal means (in which the household data used have the same value for every sampled household in a given municipality, *i.e.*, small area). These auxiliary variables are not only derived from the FIES data but also from the Philippine 2000 Labor Force Survey (LFS) and Census of Population and Housing (CPH). The LFS collects socioeconomic characteristics of the population over 15 years old. It is conducted on a quarterly basis by the NSO by personal interview, using previous week as reference period. Being part of the Integrated Survey of Households (NSCB 2000), the July 2000 and January 2001 surveys used the same sample of households as the 2000 FIES. Thus the two data sets can be merged to form a richer set of auxiliary variables. Additional auxiliary variables were also taken from the 2000 CPH in the form of municipal means. Census variables in both the short and long form were averaged at municipal level to create new data sets that could be merged with the set of auxiliary variables from FIES and LFS.

Presented in Tables 1, 2, and 3 are the computed estimates of the parameter ( $\beta$ ) and the corresponding standard errors as well as the estimates of the variance components at the national, regional and provincial levels, respectively. Table 2 is one of the regional models of the 16 models fitted at the regional level (there are 16 regions in the Philippines in the year 2000). Similarly, Table 3 shows one of the provincial models of the 20 models formulated for 20 selected provinces. To standardize comparison, exactly the same set of predictor variables are used for all the different model fitting techniques. (There are five sets of parameter estimates, although there are only four basic methods considered, because ELL is used both with and without heteroscedasticity.) Note that in practice when ELL is applied, the survey data is often subdivided and separate models fitted to each subsample, *e.g.*, to each regionally-based stratum as the 16 regions in the Philippines or even provincial level models. This can lead to overfitted models and downwardly biased standard errors for small area estimates. For the analysis here, a single model (or the national level model) has been fitted. In practice intermediate models with some but not all possible regional effects seem to work best. See for example Haslett and Jones (2005).

To assess the differences of the estimates generated from the different techniques, an informal comparison of the “significance” of the different estimates of  $\beta$  is conducted by subtracting from the estimate by one method the mean of the other methods’ estimates, then dividing by the standard error of the one method. At the national level (Table 1), estimates of the regression coefficients generated from the different methods are significantly different from each other for a number of the independent variables. GSR tends to generate estimates of the regression coefficients for the majority of the variables that are significantly different from the other methods. As pointed out earlier, the GSR estimator is the sample weighted regression estimator for a model with homoscedastic variance structure and uncorrelated observations in the population and hence this estimator is not derived under the model specified by (7). However, it is the most conservative as it generates the highest standard error for all the household level characteristics. On the other hand, the IWEE method has the highest estimated standard error for all the municipal means. The ELL\_H (ELL with heteroscedasticity) method can be considered to be the least conservative since it produces the lowest standard errors for all the estimated regression coefficients of the household level characteristics as well as for the municipal means, except for two variables where GSR generated the smallest estimates. As to the estimates of the variance components, the ELL method generates the smallest estimated cluster level variance, which is about 92% of the Pseudo-EBLUP method and 86% of the IWEE method. As to the household level variance, the IWEE method generates the smallest estimate.

**Table 1**

National level estimates of regression parameters with the standard errors and the variance components for the four techniques.  
 \*Different value for each household (mean = 0.1576633) \*\*Based from the ELL results

Explanatory Variables	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.11867	0.00181	-0.12034	0.00165	-0.11875	0.00183	-0.11888	0.00180	-0.11405	0.00216
famsizesqc	0.00937	0.00039	0.00981	0.00036	0.00938	0.00039	0.00939	0.00038	0.00898	0.00044
type_mult	0.03876	0.01697	0.03703	0.01588	0.03699	0.01717	0.03466	0.01692	0.11460	0.02194
per_kids	-0.20342	0.01476	-0.20818	0.01322	-0.20293	0.01491	-0.20216	0.01467	-0.22864	0.01617
roof_light	-0.06314	0.01291	-0.05808	0.01056	-0.06263	0.01306	-0.06175	0.01287	-0.09251	0.01413
per_61up	-0.09402	0.01420	-0.08331	0.01371	-0.09392	0.01435	-0.09389	0.01412	-0.09705	0.01698
roof_strong	0.05882	0.01135	0.05633	0.00962	0.05944	0.01148	0.06030	0.01132	0.03118	0.01293
wall_light	-0.05459	0.01182	-0.04979	0.00975	-0.05426	0.01195	-0.05392	0.01178	-0.06286	0.01353
wall_salvaged	-0.10814	0.02505	-0.11327	0.02058	-0.10748	0.02533	-0.10607	0.02495	-0.15702	0.02925
wall_strong	0.14248	0.01051	0.12964	0.00910	0.14274	0.01063	0.14319	0.01047	0.12662	0.01284
fa_xs	-0.17052	0.00941	-0.16756	0.00782	-0.17144	0.00952	-0.17236	0.00939	-0.14213	0.01110
fa_s	-0.08368	0.00861	-0.08242	0.00725	-0.08403	0.00871	-0.08454	0.00857	-0.06667	0.00964
fa_l	0.09016	0.00908	0.08478	0.00792	0.09065	0.00918	0.09106	0.00904	0.07848	0.01047
fa_xl	0.16659	0.01104	0.15404	0.00992	0.17034	0.01117	0.17121	0.01100	0.14300	0.01334
fa_xxl	0.27072	0.01144	0.24485	0.01094	0.27172	0.01157	0.27274	0.01140	0.23913	0.01457
fa_xxxl	0.36190	0.01371	0.31369	0.01286	0.36270	0.01387	0.36382	0.01367	0.32123	0.02025
all_eled	0.19084	0.01535	0.20497	0.01307	0.19031	0.01551	0.18964	0.01527	0.21344	0.01831
all_hsed	0.42325	0.01250	0.43771	0.01083	0.42192	0.01263	0.42024	0.01244	0.48180	0.01475
all_coed	1.21591	0.01371	1.29368	0.01379	1.21324	0.01386	1.20935	0.01366	1.35022	0.01827
dom_help	0.60207	0.01629	0.61218	0.01886	0.60035	0.01645	0.59733	0.01620	0.70307	0.02656
head_male	-0.05878	0.00988	-0.04581	0.00932	-0.05862	0.00998	-0.05819	0.00982	-0.07410	0.01173
no_spouse	-0.09367	0.00987	-0.07376	0.00917	-0.09361	0.00997	-0.09351	0.00981	-0.09599	0.01123
hou_9600	0.28537	0.07654	0.25643	0.07375	0.28871	0.07911	0.28783	0.08066	0.31956	0.07941
hea_rel_mus	0.09058	0.02645	0.10859	0.02507	0.09753	0.02728	0.09731	0.02782	0.10196	0.02737
Per_eng	0.17273	0.06529	0.14561	0.06298	0.17782	0.06754	0.17799	0.06887	0.17076	0.06407
Hou_coelpg	0.37463	0.04348	0.39784	0.04210	0.37934	0.04494	0.37792	0.04581	0.42682	0.03711
Hou_own_ref	0.17716	0.10497	0.18342	0.10178	0.17189	0.10843	0.17329	0.11055	0.13791	0.09766
Hou_own_tel	1.39287	0.13356	1.42109	0.12987	1.38551	0.13723	1.38974	0.13989	1.23506	0.13019
Per_wor_prh	0.46957	0.15484	0.40302	0.14926	0.47517	0.16006	0.47208	0.16317	0.50814	0.15210
Per_ind_52	-0.76245	0.21708	-0.78120	0.21073	-0.76326	0.22410	-0.76307	0.22849	-0.73294	0.21214
const	9.54013	0.05525	9.54456	0.05290	9.53566	0.05698	9.53594	0.05791	9.52622	0.05613
Variance Components Estimate	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH** level	Cluster** level
	0.18461	0.04741	NA*	0.04741	0.18820	0.05172	0.18185	0.05498	0.18461	0.04741

**Table 2**

Regional level estimates of regression parameters with the standard errors and the variance components for the four techniques.  
 \*Different value for each household (mean = 0.18930) \*\*Based from the ELL results

Explanatory Variables	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.12327	0.00760	-0.12934	0.00689	-0.12377	0.00752	-0.12380	0.00749	-0.11786	0.00997
famsizesqc	0.01096	0.00164	0.01190	0.00147	0.01101	0.00163	0.01102	0.00162	0.01030	0.00195
dom_help	0.81037	0.08873	0.75624	0.10986	0.80727	0.08784	0.80708	0.08751	0.84490	0.08911
wall_light	-0.06808	0.04289	-0.06390	0.03743	-0.06020	0.04272	-0.05973	0.04257	-0.14472	0.04226
wall_strong	0.13761	0.03745	0.15212	0.03469	0.14514	0.03737	0.14560	0.03725	0.06116	0.04249
fa_xs	-0.22074	0.04910	-0.22368	0.04518	-0.22723	0.04875	-0.22761	0.04858	-0.14856	0.05665
fa_s	-0.13540	0.03840	-0.12255	0.03344	-0.13775	0.03805	-0.13789	0.03791	-0.11059	0.04538
fa_l	0.09484	0.03709	0.08894	0.03429	0.09590	0.03676	0.09597	0.03663	0.08529	0.04122
fa_xl	0.16627	0.04315	0.15519	0.04072	0.16938	0.04284	0.16958	0.04269	0.13698	0.04897
fa_xxl	0.33706	0.04545	0.31196	0.04829	0.34173	0.04516	0.34201	0.04500	0.29156	0.05148
fa_xxxl	0.33103	0.06185	0.30377	0.06029	0.33762	0.06134	0.33801	0.06111	0.26052	0.06635
all_hsed	0.33987	0.05253	0.35591	0.04783	0.33807	0.05209	0.33796	0.05189	0.35776	0.04843
all_coed	1.21824	0.05734	1.24762	0.05842	1.20787	0.05692	1.20726	0.05671	1.32979	0.06227
per_kids	-0.24699	0.06440	-0.24047	0.05846	-0.24439	0.06371	-0.24424	0.06347	-0.27423	0.07050
per_61up	-0.14609	0.06126	-0.15938	0.05787	-0.14703	0.06063	-0.14708	0.06040	-0.13525	0.07124
hou_9600	1.13985	0.49103	1.27035	0.47888	1.14320	0.52137	1.14357	0.52172	1.07509	0.51937
Hou_own_ref	1.45233	0.24550	1.51020	0.23864	1.44986	0.26072	1.44985	0.26089	1.44779	0.23585
const	9.36877	0.20322	9.32363	0.19660	9.36597	0.21502	9.36569	0.21512	9.41385	0.21430
Variance Components Estimate	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH** level	Cluster** level
	0.19544	0.03073	NA*	0.03073	0.19052	0.03728	0.18902	0.03748	0.19544	0.03073



**Table 3**  
Provincial level estimates of regression parameters with the standard errors and the variance components for the four techniques.  
\*Different value for each household (mean = 0.23749) \*\*Based from the ELL results

Explanatory Variables	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1450	0.0175	-0.1489	0.0156	-0.1452	0.0179	-0.1449	0.0171	-0.1413	0.0097
famsizesqc	0.0090	0.0063	0.0124	0.0067	0.0091	0.0065	0.0090	0.0062	0.0085	0.0055
fa_xs	-0.4549	0.1126	-0.3816	0.1010	-0.4552	0.1149	-0.4546	0.1095	-0.4479	0.0718
fa_s	-0.2550	0.0976	-0.2653	0.0794	-0.2545	0.0995	-0.2555	0.0951	-0.2693	0.1198
wall_light	-0.2055	0.0945	-0.1474	0.0778	-0.2057	0.0965	-0.2058	0.0919	-0.2063	0.1070
all_hsed	0.4007	0.1643	0.3531	0.1448	0.4015	0.1673	0.4006	0.1601	0.3891	0.1585
all_coed	1.5411	0.1677	1.8202	0.1769	1.5429	0.1709	1.5429	0.1635	1.5439	0.2326
Hou_own_tel	3.4373	1.0270	3.2630	1.0582	3.4265	1.0622	3.4274	0.9871	3.4392	0.5733
Per_wor_prh	-1.1075	1.1933	-1.5801	1.2008	-1.1049	1.2327	-1.1056	1.1483	-1.1150	0.8729
const	10.0976	0.1480	10.0798	0.1279	10.0988	0.1517	10.0981	0.1435	10.0872	0.1373
Variance Components Estimate	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH** level	Cluster** level
	0.25753	0.01871	NA*	0.25753	0.26682	0.02079	0.24498	0.01671	0.25753	0.01871

At the regional level, estimates of the regression coefficients are generally similar for all the different estimation methods, except that the GSR and/or ELL\_H methods generated estimates for a few variables which were significantly different from the other methods. Similar to the national level estimated standard errors, GSR also tends to be the most conservative method for the majority of the regional level models - it generated the highest estimated standard errors for most of the regression coefficients of the household characteristics. IWEE has the highest estimated standard error for most of the coefficients of the municipal means. The ELL\_H method produces the lowest standard errors for the majority of the regression coefficients of the household characteristics and municipal means. The ELL method tends to generate the smallest estimated cluster level variance with ratios to Pseudo-EBLUP and IWEE ranging from around 82% to 100%. The IWEE method still has the smallest household level variance.

Similar to the regional level estimates, the regression coefficients' estimates at the provincial level are similar except for some discrepancies from the GSR and ELL\_H estimates. For the estimated standard errors of the regression coefficients, the ELL\_H still produces the lowest estimates for the majority of the coefficients of the household characteristics; however, the GSR method (instead of the ELL\_H method) now produces the lowest estimated standard error for the majority of the municipal means. The ELL method still tends to generate the smallest estimated cluster level variance for most provinces with the smallest ratio to Pseudo-EBLUP about 53% and to IWEE about 48%. For a number of provinces, IWEE tends to generate the smallest estimated cluster level variance. For the household level variance, IWEE still generated the smallest estimate. Generally, estimates of the cluster level variance tend to be more variable at the provincial level which is due to smaller sample sizes.

For small area estimates of poverty, after the regression model is applied to census data, estimated standard errors in

the regression are only one part of the small area estimates' standard errors. There is also variation at the cluster level in (7) that needs to be considered (to different degrees depending on the level of aggregation used to construct the small areas) and there is variation at household level too. These additional sources of variation can be assessed via the estimated variance components. As shown above, regardless of the level (national, regional and provincial) at which the model is formulated, the IWEE method generates the smallest household level variance, while the ELL method generates the smallest cluster level variance. Since the cluster level variation usually makes a much larger contribution to the estimated standard error at the small area level, ELL is again the least conservative. We note that the household level variance under the ELL method with heteroscedasticity model varies from one unit to another, hence, the mean value is reported, and that the estimated  $R^2$  for the heteroscedasticity model is negligible,  $R^2 = 0.03$  even at the national level, so that in terms of regression model fit at least it may offer few advantages for this data set. In our experience with applying the ELL method we have found that heteroscedasticity modeling is unnecessary.

Returning to the regression (*i.e.*, the estimates generated for  $\beta$  and the estimated standard error for the different techniques), IWEE is the method that best incorporates the sampling weights from the computation of the variance components necessary for the generation of small area estimates and their estimated standard errors. In terms of implementation, the GSR method would generally be the simplest option as it is available for example in packages such as Stata, Sudaan or WesVar. The ELL method combines sampling weights and covariance structure in a way that is non-standard in that it uses an estimate of  $\mathbf{W}_b \mathbf{V}_b^{-1}$  in (8) and (9) to produce an asymmetric estimated covariance matrix for the estimates of  $\beta$  and for estimating  $\beta$  itself. For estimating  $\beta$  this would be acceptable if the asymmetric matrix were a generalized inverse of the correct covariance matrix. It is however clearly not acceptable as an



estimated covariance matrix, a problem ELL attempt to circumvent (*e.g.*, in the World Bank's POVMAP software) by averaging each of the relevant pairs of off-diagonal elements to meet the necessary condition that a covariance matrix be symmetric.

Generally in the ELL method of poverty estimation only variables matching in terms of average and standard deviation in both survey and census plus census averages can be used. This is because, after the regression model has been fitted to the survey data, in the second phase it is applied to the census data as a predictor at household level, *i.e.*, the regression equation (however it has been estimated) is used to find predicted values of per capita income or expenditure for each census household, generated via

$$\hat{Y}_{bh} = \mathbf{x}'_{bh} \hat{\beta} + \hat{v}_b + \hat{e}_{bh} \quad (25)$$

using imputed values of  $v_b$  and  $e_{bh}$  (based for example on bootstrap sampling from their survey estimates). Here  $\mathbf{x}_{bh}$  are auxiliary variables from the census. Poverty indices are typically based on non-linear functions of log-income or log-expenditure, so the predictions from (25) are transformed appropriately before averaging over each small area. Note that in practice  $v_b$  can be estimated for the sampled clusters, but the sample and census codes usually do not match so these cannot be identified in the census, and it is the bootstrap (by selecting from the sampled barangays, *i.e.*, PSUs) that provides imputed values for all barangays; a parallel comment applies to  $\hat{e}_{bh}$  for households within clusters. The general benefit of using census data in this way (as ELL does) is that the predictor variables can be used for all census households (of which there are many) not just those in the survey, thereby increasing accuracy of the small area estimates (conditional on the model being correct). Note that the estimates in (25) remain unbiased even if  $v_b$  and  $e_{bh}$  are not included in the prediction itself, but the variance estimate for small area  $a$  needs to be computed based on equation (25) so that it incorporates the necessary additional variation at cluster and household levels.

In poverty estimation, we are interested in area-level summaries of non-linear functions of  $\hat{Y}_{bh}$ , for example, whether it is below the poverty line (poverty incidence) and poverty gap rather than the regression fitting per se. It is instructive here to examine the effects of model uncertainty on area mean estimates

$$\bar{y}_a = \bar{\mathbf{x}}'_a \hat{\beta} \quad (26)$$

where  $\bar{\mathbf{x}}_a$  is the population (*i.e.*, census) mean for area  $a$  of the covariates including the constant 1, after the regression model has been applied to the census data as in phase 2 of ELL. By similarly averaging (7) to get the true mean  $\bar{Y}_a$ , subtracting from (26), and applying the variance operator, we get the prediction error variance equation:

$$V(\bar{y}_a - \bar{Y}_a) = \bar{\mathbf{x}}'_a \Phi_w \bar{\mathbf{x}}_a + \frac{1}{N_a^2} \sum_{b=1}^m N_b^2 \sigma_v^2 + \frac{1}{N_a} \sigma_e^2 \quad (27)$$

where  $N_a$  is the population size at a particular level of aggregation,  $N_b$  is the population size in each cluster,  $\Phi_w$  is the variance-covariance matrix of the regression coefficient estimates, and  $(\sigma_v^2, \sigma_e^2)$  are the cluster and household level variance components, respectively. Note that estimating this prediction error variance requires estimates of the variance components, but any bias caused by uncertainty in these would be a second order effect (see Prasad and Rao 1990).

Based on (27), the extent of the influence of the survey based regression model and other variance components (cluster and household level) on the accuracy of the final small area estimates can be compared for any fitting technique and/or levels of aggregation. Generally, it is either the regression model (via the estimate of the regression parameters) or the cluster effect that dominates the estimated accuracy of the computed small area estimate. Using the national level model in Table 1 and the survey data (instead of the census) auxiliary variables to estimate the first term in (27), shows that the extent to which the regression model effect contributes to small area estimate variance increases markedly as household data are more aggregated - about 0.25% at the municipal level, 20% at the provincial level and 70% at the regional level. In other words, the more aggregated the data into larger areas, the greater the dominance of the regression model parameter uncertainty, regardless of the regression fitting method. This is as expected because even at high levels of aggregation, the contribution to the overall variance from the model effect depends on the average covariate values, not on the population size. This is the reason that, at the most aggregated regional level, small area techniques usually offer little improvement over direct estimates. This is also why it is important (as this paper has done) to examine in detail the regression fitting procedures applied in small area estimation of third world poverty.

The effect of cluster level variation is different: at lower levels of aggregation (*e.g.*, municipality) the computed variance of the small area estimates are dominated by the cluster component of variance or cluster level effect, *i.e.*, for small areas (other than regional estimates) the variance component, not the regression model, has the greatest impact on the value of the standard error of the small area estimates. Consequently, the accuracy of estimates of variance components especially at cluster level can be crucial to accurate estimation of standard error of small area estimates at the aggregation level at which they are most useful (for example at municipal level in the Philippines). Again, this is why the method used for phase 1 fitting for variance components as discussed in this paper, are critical to small area estimation of poverty.

Presented in Tables 4-6 are Kruskal-Wallis (KW) tests (Siegel 1956) for the various fitting methods conducted on the estimated variances at the municipal (Table 4), provincial (Table 5) and regional (Tables 6) levels. In Table 4 significant differences exist among the variance estimates generated by the various small area techniques, as shown by the p-values of the Kruskal-Wallis statistics. Multiple comparison of mean ranks shows the Pseudo-EBLUP and IWEE methods have variance estimates at cluster level that are significantly higher than the other methods, but not significantly different from each other (although for the IWEE method the Z-value for the difference from average rank is in general rather higher than all the others).

The ELL method and the GSR method generate significantly lower and similar variance component estimates. This is principally because we used the ELL variance components estimation technique in generating variance components for the GSR method (because GSR does not usually estimate variance components), although the residuals we used were not identical for the two regression fitting methods. As expected, at the municipal level for which small area estimates were used in practice, the cluster effect (rather than regression coefficient uncertainty) is generally the dominant part of the small area variance estimates. Since the ELL and GSR methods have similar cluster level variance, their corresponding variance estimates at small area also tend to be similar. Explicitly, observe from Table 4 that the ranking of the variance estimates generally conforms with the ranking of the cluster effects.

In poverty estimation, estimates at higher levels of aggregation, such as those in Table 5 and 6, are generally carried out for comparison with direct survey estimates at these more aggregated levels, even though they are not particularly useful for aid allocation. The results do however, support those indicated for lower level of aggregation. In Table 5 and Table 6, the estimated variances for the poverty estimates generated by the different techniques are not significantly different from each other at the provincial and regional level, an effect that is partially due to the small number of provinces and even smaller number of regions. The variances and hence the standard errors may not be significantly different from each other, but it is worth noting that the GSR method tends to generate the smallest estimated standard error for the regression model and in turn the smallest variance estimate for poverty at the regional level, even though GSR generates higher standard errors for the individual regression coefficients (corresponding to the diagonal elements only in the estimated covariance matrix of  $\hat{\beta}$ ). As expected, at an even higher level of aggregation for all methods, the relative effect of the regression component is more pronounced.

The general conclusion is that, whether fitting survey data alone or using survey based regression parameter estimates in conjunction with census data, it is crucial not only to find a suitable model (*i.e.*, set of regressors) based on an adequate sample size, but also to get sound estimates of the regression parameters and their standard errors under this model as well as good estimates of the variance components at all relevant levels of aggregation. Usually the relevant levels of aggregation are determined via the survey design, rather than simply through the level at which small area estimates are sought, although the number of levels need not be limited to two (*e.g.*, to cluster-level and household-level).

Survey data, whether used for poverty estimation or in other context, also introduces problems involving survey weights that can be important not only for regression parameter estimation (and their estimated standard errors) but also for estimating variance components. Incorporating survey weights into regression models with correlated data introduces problems because it is the population correlation as it applies to the weighted survey data that needs to be properly modeled, so that weighting correlation matrices using matrix multiplication (as ELL do) is not technically adequate (see Appendix).

For the Philippine data and for the specified list of regressors, regardless of which of the four methods are used, parameter estimates were very similar, which suggests that the more important issue is possible underestimation of standard errors of parameter estimates and of variance components particularly at cluster level. ELL is the least conservative in that it gave the lowest estimates of both variance measures, and in this respect (as with its use of asymmetric estimated covariance matrices) some caution may be warranted with the regression and variance component aspects of the ELL technique. GSR gave similar estimates of standard errors for the small area estimates to ELL when using the same technique for variance components, despite having higher standard errors (and using a sound covariance matrix) for regression parameters. This is because when there is less aggregation, the level at which most small area estimates are actually used, variance components dominate.

The Pseudo-EBLUP and IWEE methods incorporate survey weights correctly (given a suitable choice of pseudo-likelihood and hence GEE) and gave larger (*i.e.*, more conservative) estimates of cluster level variance components. This suggests that these two methods and particularly IWEE are among the best of the currently available methods, not necessarily for estimating regression equations (where availability of standard software may give GSR an advantage), but for estimating the crucial variance components.



**Table 4**  
Kruskal-Wallis test for estimated variances at the municipal level (N = 1,243)

SAE	Cluster Effect			Beta Effect			Variance		
	Median	Mean Rank	Z	Median	Mean Rank	Z	Median	Mean Rank	Z
ELL(no hetero)	0.002843	2,961.2(a)	-3.22	0.0002311	3,067.3(ab)	-0.89	0.00318	2,963.4(a)	-3.18
ELL(w/ hetero)	0.002843	2,961.2(a)	-3.22	0.0002128	2,802.0(c)	-6.72	0.00316	2,930.8(a)	-3.89
Pseudo-EBLUP	0.003094	3,229.4(b)	2.67	0.0002449	3,257.5(ad)	3.28	0.00346	3,241.3(b)	2.93
IWEE	0.003294	3,426.9(b)	7.01	0.0002529	3,364.5(d)	5.64	0.00366	3,441.3(b)	7.32
GSR(Stata)	0.002843	2,961.2(a)	-3.22	0.0002311	3,048.7(b)	-1.3	0.00317	2,963.1(a)	-3.18
Overall		3,108			3,108			3,108	
KW Statistic	H = 69.92	(P = 0.000)		H = 72.19	(P = 0.000)		H = 78.06	(P = 0.000)	

**Table 5**  
Kruskal-Wallis test for estimated variances at the provincial level (N = 83)

SAE	Cluster Effect			Beta Effect			Variance		
	Median	Mean Rank	Z	Median	Mean Rank	Z	Median	Mean Rank	Z
ELL(no hetero)	0.0002518	200.3	-0.65	0.0001162	207.7	-0.03	0.00039	202.3	-0.48
ELL(w/ hetero)	0.0002518	200.3	-0.65	0.0001095	190.1	-1.52	0.00038	196.3	-0.99
Pseudo-EBLUP	0.000274	214.9	0.59	0.0001239	224.2	1.37	0.00042	217.1	0.78
IWEE	0.0002916	224.2	1.38	0.0001287	234.1	2.22	0.00045	227.8	1.68
GSR(Stata)	0.0002517	200.3	-0.65	0.00010	184	-2.04	0.00037	196.4	-0.98
Overall		208			208			208	
KW Statistic	H = 2.82	(P = 0.589)		H = 10.61	(P = 0.031)		H = 4.48	(P = 0.344)	

**Table 6**  
Kruskal-Wallis test for estimated variances at the regional level (N = 16)

SAE	Cluster Effect			Beta Effect			Variance		
	Median	Mean Rank	Z	Median	Mean Rank	Z	Median	Mean Rank	Z
ELL(no hetero)	0.000050	38.2	-0.45	0.000077	40.9	0.08	0.00013	39.3	-0.23
ELL(w/ hetero)	0.000050	38.2	-0.45	0.000073	35.1	-1.05	0.00012	37	-0.67
Pseudo-EBLUP	0.000055	42.6	0.4	0.000082	46.9	1.23	0.00014	44	0.67
IWEE	0.000058	45.3	0.93	0.000085	50.1	1.85	0.00015	46.6	1.17
GSR(Stata)	0.000050	38.2	-0.45	0.000070	29.6	-2.1	0.00013	35.6	-0.94
Overall		40.5			40.5			40.5	
KW Statistic	H = 1.30	(P = 0.861)		H = 8.36	(P = 0.079)		H = 2.58	(P = 0.630)	

Of course, such considerations (while central) need to be predicated by adequate data cleaning, sound matching of possible regressor variables (in terms of mean, variance, and meaning) between survey and census where census data is also being used. Also needed are the proper, time consuming consideration of a wide range of possible regressor variables and recognition of the limits placed on subdividing survey data by small sample sizes, since all estimated standard errors for both regression parameter and small area estimates (whatever method is used for fitting the variance component estimate) are conditional on the regression model being correct.

## 8. Conclusion and recommendation

There is a great need for sound poverty statistics in order to effectively monitor interventions and assistance to various impoverished localities. Small area estimation techniques are one methodology that is being used to provide such statistics. In this sense the issues raised in this paper concerning the accuracy of the small area estimates are not simply an academic issue but are central to the Millennium

Development Goals and to aid allocation in what is a multi-billion dollar industry.

In this paper, we have considered four estimation techniques for fitting regression models using survey data and related them to small area poverty estimation. We have shown that although differences in estimates are insufficient to invalidate the published national studies, the most frequently implemented survey data fitting technique, ELL with heteroscedasticity, recommended by the World Bank, has some limitations since (like its homoscedastic version) it lacks sound theoretical underpinning. Replacing the survey fitting part of the ELL method is recommended. For the other methodologies considered (the Pseudo-EBLUP, IWEE, and the GSR method), all have valid theoretical basis mathematically and the results generated can be clearly interpreted once the assumptions have been checked. The different methodologies when applied to complex weighted survey data from the Philippines indicate that for variance component estimation from survey data and hence for small area estimation at a fine level, Pseudo-EBLUP and particularly IWEE are likely to be better than the GSR or the ELL methods, although GSR is sound and easy to use because it is available in off-the-shelf software.



We have also shown that at the level where small area estimation is actually used for aid allocation, the variance estimate of the small area tends to be dominated by the cluster level variance rather than by the accuracy of the regression parameter estimates. Hence, it is particularly important that the cluster-level component of variance (and, if fitted as recommended, any small area level variance component) is properly estimated. It is also important that the regression model used in the generation of small area estimates (including choice of suitable regressors) is appropriate. Essentially, at lower levels of aggregation it is the variance components that dominate the standard error of the small area estimates, so that the estimation of the variance components is critical whatever the choice of aggregation level. Sound survey-based regression method, good choice of regression variables, and care with sample size (especially if separate regression models are fitted to subsets of survey data), also remain central to sound small area estimation of third world poverty.

### Acknowledgements

The authors would like thank the referees and the Associate Editor for their careful reading of the manuscript and for their helpful suggestions.

### Appendix

In footnote 8 of the Elbers *et al.* (2002) World Bank working paper and implicitly in Elbers *et al.* (2003) in *Econometrica*, the covariance of the error process is denoted  $\Omega$  and it is stated that  $\mathbf{W}\Omega^{-1} = \mathbf{P}^T\mathbf{P}$  where  $\mathbf{W}$  is 'a weighting matrix of expansion factors'. In the notation of Section 4 above,  $\mathbf{W}$  is block diagonal with or diagonal with diagonal blocks  $\mathbf{W}_b$ , and  $\Omega$  is block diagonal with diagonal blocks  $\mathbf{V}_b$ .

However, either  $\mathbf{W}$  and  $\Omega$  (or  $\Omega^{-1}$ ) are non-conformable (with weighting factors in  $\mathbf{W}$  at cluster level and the observations and hence  $\Omega^{-1}$  at individual level), or if conformable  $\mathbf{W}\Omega^{-1}$  is generally asymmetric (even if  $\mathbf{W}$  is diagonal) unless  $\mathbf{W}$  is a simple multiple of the identity matrix, *i.e.*,  $\mathbf{W} = \sigma^2\mathbf{I}$ .

Hence,  $\mathbf{W}\Omega^{-1}$  does not equal  $\mathbf{P}^T\mathbf{P}$  as has been claimed since  $\mathbf{P}^T\mathbf{P}$  is symmetric in general and  $\mathbf{W}\Omega^{-1}$  is not. Making  $\mathbf{W}\Omega^{-1}$  symmetric by adding it to its transpose and dividing by two, as is done in the World Bank PovMap software, is not a technically adequate solution to this problem. (Note that even in the simple case where  $\mathbf{W}$  and  $\Omega^{-1}$  are conformable, and  $\mathbf{W}$  is diagonal but not all diagonal elements are equal,  $\mathbf{W}\Omega^{-1}$  is not diagonal because it has every element of row  $i$  of  $\Omega^{-1}$  multiplied by  $w_i$

(where  $w_i$  is the  $i^{\text{th}}$  diagonal element of  $\mathbf{W}$ ) but the  $i^{\text{th}}$  column does *not* have every element multiplied by an identical weight.)

Putting this issue of symmetry to one side, and using  $\mathbf{P}^T\mathbf{P}$  in place of  $\mathbf{W}\Omega^{-1}$ , ELL seem to be claiming that comparing their 'sample survey adjusted weighted GLS estimator' to the 'unadjusted GLS' estimator implies that instead of using  $\Omega^{-1}$  as the underlying metric (*i.e.*, the inverse of the relevant covariance matrix), a weighted version namely  $\mathbf{W}\Omega^{-1}\mathbf{W}^T$  should be used. This creates no asymmetry issue in itself (provided  $\mathbf{P}^T\mathbf{P}$  were used in place of  $\mathbf{W}\Omega^{-1}$ ). However, even if  $\mathbf{W}$  were diagonal and  $\mathbf{P}^T\mathbf{P}$  used, the weight matrix  $\mathbf{W}$  cannot use even unequal diagonal weights corresponding to the sampled units, *i.e.*,  $w_i$  say, because the  $ij^{\text{th}}$  element of  $\Omega^{-1}$  (unlike the  $ij^{\text{th}}$  element of  $\Omega$ ) does *not* correspond to the  $i^{\text{th}}$  and  $j^{\text{th}}$  unit in the sample (or in the population), so it is rather unclear what  $\mathbf{W}$  is or how  $\mathbf{W}$  can be sensibly defined as 'a weighting matrix of expansion factors'.

This argument still applies when  $\mathbf{V}_b$  is replaced by its estimator  $\hat{\mathbf{V}}_b$  which uses estimates in place of  $\sigma_e^2$  and  $\sigma_v^2$ .

### References

- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- Chambers, R. (2006). What is poverty? Who asks? Who answers? *Poverty in Focus*, UNDP, December 2006, 3-4.
- Elbers, C., Lanjouw, J. and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71, 355-364.
- Elbers, C., Lanjouw, J. and Lanjouw, P. (2002). *Micro-level Estimation of Welfare*. Research Working Paper 2911, World Bank, Development Research Group, Washington, D.C.
- Ghosh, M., and Rao J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- Haslett, S., and Jones, G. (2004). *Local Estimation of Poverty and Malnutrition in Bangladesh*, Bangladesh Bureau of Statistics and United Nations World Food Programme.
- Haslett, S., and Jones, G. (2005). *Local Estimation of Poverty in the Philippines*, Philippine National Statistics Co-ordination Board/World Bank Report. [http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Local\\_Estimation\\_of\\_Poverty\\_Philippines.pdf](http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Local_Estimation_of_Poverty_Philippines.pdf).
- Haslett, S., and Jones, G. (2005). Small area estimation using surveys and censuses: Some practical and statistical issues. *Statistics in Transition*, 7, 541-556.
- Haslett, S., and Jones, G. (2006). *Small Area Estimation of Poverty, Caloric Intake and Malnutrition in Nepal*. Published: Nepal Central Bureau of Statistics/World Food Programme, United Nations/World Bank, September 2006, 184pp, ISBN 999337018-5.

- Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226-252.
- Horton, N.J., and Lipsitz, S.R. (1999). Review of software to fit Generalized Estimating Equation regression models. *The American Statistician*, 53, 160-169.
- Liang, K.L., and Zeger, S. (1986). Longitudinal data analysis using Generalized Linear Models. *Biometrika*, 73, 13-22.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Brooks/Cole Publishing Company.
- Militino, A.F., Ugarte, M.D., Goicoa, T. and Gonzalez-Audicana, M. (2006). Using small area models to estimate the total area occupied by olive trees. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 450-461.
- Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B*, 60, 23-40.
- Pfeffermann, D., Moura, F.A. and Silva, P.L. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 949-959.
- Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.
- Rao, J.N.K. (2003). *Small Area Estimation*, Wiley Series in Survey Methodology. Wiley-Interscience, John Wiley & Sons, Inc.
- NSCB (2000). *Profile of Censuses and Surveys*. National Statistical Coordination Board, Philippines.
- Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Series in Psychology. New York: McGraw-Hill.
- Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Chichester: John Wiley & Sons.
- UN website. <http://www.un.org/millenniumgoals/>.
- You, Y., and Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.
- You, Y., Rao, J.N.K. and Kovačević, M. (2003). Estimating fixed effects and variance components in a random intercept model using survey data. *Proceedings: Symposium 2003, Challenges in Survey Taking for the Next Decade*. Statistics Canada.
- Zhao, Q. (2006). User manual for PovMap, The World Bank. [http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Zhao\\_ManualPovMap.pdf](http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Zhao_ManualPovMap.pdf).

## Small area estimation of the number of firms' recruits by using multivariate models for count data

Maria Rosaria Ferrante and Carlo Trivisano<sup>1</sup>

### Abstract

The number of people recruited by firms in Local Labour Market Areas provides an important indicator of the reorganisation of the local productive processes. In Italy, this parameter can be estimated using the information collected in the Excelsior survey, although it does not provide reliable estimates for the domains of interest. In this paper we propose a multivariate small area estimation approach for count data based on the Multivariate Poisson-Log Normal distribution. This approach will be used to estimate the number of firm recruits both replacing departing employees and filling new positions. In the small area estimation framework, it is customary to assume that sampling variances and covariances are known. However, both they and the direct point estimates suffer from instability. Due to the rare nature of the phenomenon we are analysing, counts in some domains are equal to zero, and this produces estimates of sampling error covariances equal to zero. To account for the extra variability due to the estimated sampling covariance matrix, and to deal with the problem of unreasonable estimated variances and covariances in some domains, we propose an "integrated" approach where we jointly model the parameters of interest and the sampling error covariance matrices. We suggest a solution based again on the Poisson-Log Normal distribution to smooth variances and covariances. The results we obtain are encouraging: the proposed small area estimation model shows a better fit when compared to the Multivariate Normal-Normal (MNN) small area model, and it allows for a non-negligible increase in efficiency.

Key Words: Multivariate Poisson-Log Normal distribution; Zero counts; Generalized Variance Function; Hierarchical Bayesian models.

### 1. Introduction

The number of people recruited by firms for a certain period can be taken as a key indicator of ongoing changes in the economic system. To highlight the dynamic of the demand for local labour, we consider the number of people recruited by firms in Local Labour Market Areas (LLMAs), these last grouped according to i) productive specialization, ii) firms' size classes and iii) industrial sector. Domains are defined by cross-classifying these three variables. In order to emphasise the signals of the reorganisation of the productive process, we focus on the numbers of "recruits replacing employees leaving the firm (substitute recruits – SR)" and "recruits filling new positions (new recruits – NR)". In Italy, information about firms' recruits is collected by the Excelsior Survey co-sponsored by the Union of Italian Chambers of Commerce (UNIONCAMERE), the Ministry of Labour and the European Union. Unfortunately, this survey does not provide reliable estimates of firms' recruits for each of these domains due to small domain sample size. As a consequence, a small area estimation (SAE) technique has to be adopted in order to obtain estimates with an acceptable degree of variability.

In this paper, we propose a SAE approach for the estimation of counts. Due to data constraints, we adopt an aggregated area-level model.

Since we aim at estimating SR and NR, we adopt a multivariate SAE model that borrows strength not only from areas but also from the correlations between the NR and SR true values. In order to estimate the median income of different sized groups of families, Fay (1987) proposed a multivariate regression model in an Empirical Bayes context. Multivariate SAE approaches have also been developed by Ghosh, Nangia and Kim (1996) and Datta, Fay and Ghosh (1991), Datta, Ghosh, Nangia and Natarajan (1996) and Datta, Lahiri, Maiti and Lu (1999) for continuous data in the hierarchical cross-section time series model framework. Fabrizi, Ferrante and Pacei (2005, 2008) adopted multivariate area level models to estimate a vector of continuous poverty parameters. As in the univariate Fay-Herriot model (Fay and Herriot 1979), all of the papers mentioned above assume the use of small area normal sampling and linking models.

Since the sampling correlations between SR and NR estimators are mainly negative, we propose a SAE model based on the Multivariate Poisson-Log Normal (MPLN) distribution. Unlike other multivariate distributions for counts proposed in the literature, this particular distribution allows for unconstrained (that is, both positive and negative) correlations (Aitchison and Ho 1989).

We also deal with the instability of estimators of sampling error variances and covariances. An approximately unbiased estimate of the variance of direct estimators is

1. Maria Rosaria Ferrante, Department of Statistics - University of Bologna, Italy. E-mail: maria.ferrante@unibo.it; Carlo Trivisano, Department of Statistics - University of Bologna, Italy. E-mail: carlo.trivisano@unibo.it.



usually available in SAE. However, in area-level models it is customary to assume that the sampling variance is known and equal to its estimate (Rao 2003; page 76). This assumption is commonly stated and largely accepted in the case of large samples, whereas both the variance estimator and direct point estimators suffer from instability in the case of small samples. As a partial solution, sampling variance estimates are often smoothed through the generalized variance functions (GVF) approach (Wolter 1985). In You, Rao and Gambino (2003), sampling variances and covariances were smoothed over areas and times. In order to consider the extra variability associated with the estimated sampling variances, Arora and Lahiri (1997) proposed an integrated Hierarchical Bayes (HB) smoothing approach for continuous data. See You and Chapman (2006), Liu, Lahiri and Kalton (2007) and You (2008) for different extensions of Arora and Lahiri (1997).

Due to the rarity of recruits in certain domains, a further problem arises that is linked to the instability of sampling error variances and covariances estimators. When direct estimates of SR or NR (or both) are equal to zero, estimated sampling error variances and covariances are also equal to zero. Note that observing estimated variances equal to zero does not necessarily imply that the estimates have a high degree of accuracy. This problem was encountered in previous small area estimation problems (*e.g.*, Elazar 2004; Chattopadhyay, Lahiri, Larsen and Reimnitz 1999). Chen (2001) proposed a unit level hierarchical modeling to handle the problem. Moreover, some studies (Cohen 2000) use the logarithmic transformation of the mean (or total) direct estimates of the count data in order to adopt a linear SAE model, simply discarding the estimates equal to zero. Although this solution overcomes the "zero variance" problem, it also leads to biased estimates and neglects a portion of the sample.

In order to deal with the instability of variances and covariances estimators as well as the problem of estimated sampling variances equal to zero, we suggest an "integrated" approach in the spirit of that proposed by Arora and Lahiri (1997), Liu *et al.* (2007) and You (2008). Within an HB framework, we jointly model the parameters of interest and the sampling error covariance matrices by adopting a smoothing covariance solution based once again on the Poisson-Log Normal distribution.

The layout of this paper is as follows. The data set employed is described in section 2, while section 3 presents direct domain estimation and its associated sampling error variances and covariances. In section 4, we describe the multivariate SAE model we propose for estimating counts as well as the solution we suggest for overcoming the instability of sampling error variances and covariances estimators in the presence of zero counts. Section 5 reports

the results obtained by measuring the performance of the adopted SAE model. Details on the Poisson-Log Normal distribution are given in the Appendix.

## 2. The excelsior survey

The Excelsior Survey is one of the most complete Italian statistical sources for labour demand data, providing estimates of the number of people recruited by Italian firms. Each year, a stratified simple random sample of about 100,000 firms with at least one employee is contacted and asked about the number of people it plans to hire in the short term. The factors used for stratification are the firm's industrial sector and size class. The allocation of the sample in the strata satisfies a constraint on the maximum estimated standard error corresponding to a 95% significance level (Baldi, Bellisai, Fivizzani and Sorrentino 2007). By focusing on local geographical details, the survey is designed to produce reliable estimates for the administrative provinces (NUTS3, following the "Nomenclature of Units for Territorial Statistics" reported in <http://europa.eu.int/comm/eurostat/ramon/nuts>). This geographical unit, singled out on the basis of administrative criteria, does not appear to be the best choice when analysing the dynamics of the local labour demand. In order to shed some light on the signals of the reorganization of the local productive process, a better territorial subdivision would be LLMAAs (following the OECD definition). LLMAAs are groups of municipalities sharing the same labour market conditions (for the location of LLMAAs in Italy, see Sforzi 1991). In Italy, following the strategy proposed by Sforzi and Lorenzini (2002) and adopted by the Italian Statistical Institute (ISTAT), certain LLMAAs are labelled "industrial districts" (IDs). IDs are geographically defined productive systems characterized by a dominant specialization. In the 1990s, these were considered to be the main stimulus for the growth of the Italian economic system (Becattini 1992).

Estimating the number of substitute and new recruits in firms operating within/outside of IDs can help us verify whether IDs are still a source of dynamism for the Italian economy as a whole. In order to refer to types of ID, we group them according to their productive specialization. Similarly, LLMAAs not labelled as IDs can be classified according to their economic vocation (LLMAAs can be characterized by a specific manufacturing activity, tourist area, city, *etc.*). Moreover, the comparison between ID and non-ID firms makes economic sense if the industrial sector and size of the firms are also taken into account. Finally, as already noted, domains of interest are defined by cross-classifying: i) groups of LLMAAs obtained according to their productive specialization, ii) firm's industrial sector and iii) firm's size.

This paper focuses on the manufacturing sector characterising the IDs' economic activity. The analysis is limited to two Italian regions containing a large quantity of IDs, namely Tuscany and Emilia-Romagna, and to firms with fewer than 100 employees (as censuses are taken for the other size classes). The target population consists of 54,089 firms employing a total of 809,059 people.

### 3. Direct estimates

Table 1 provides details of the categories defining the 208 domains of interest. Note that the number of domains is less than that expected due to the absence of a number of domains within the population. The domains are unplanned since they are formed grouping LLMAAs contained in the same planned stratum. For the sake of simplicity, in the following we avoid using the stratum subscription wherever possible.

Let  $\theta_{i1}$  and  $\theta_{i2}$  be the true number of NR and SR for domain  $i$  ( $i = 1, \dots, 208$ ), respectively. We shall first define a direct estimator of  $\theta_{ij}$  ( $i = 1, \dots, 208$ ;  $j = 1, 2$ ). Let  $y_{ijl}$  be the response of the  $l^{\text{th}}$  unit related to the  $j^{\text{th}}$  variable in the  $i^{\text{th}}$  domain ( $l = 1, \dots, n_i$ , where  $n_i$  is the sample size in domain  $i$ ;  $i = 1, \dots, 208$ ;  $j = 1, 2$ ). As design based (direct) estimator we use a ratio domain estimator defined as  $\hat{\theta}_{ij} = \sum_{l=1}^{n_i} y_{ijl} / (n_i / N_i) N_i / \hat{N}_i$ , where  $N_i$  and  $n_i$  are respectively the population size and the sampling size referred to domain  $i$ , and  $\hat{N}_i = n_i / n_{i\pi} N_{i\pi}$ , where  $N_{i\pi}$  and  $n_{i\pi}$  are respectively the population size and the sampling size of the

stratum  $t$  containing the domain  $i$  (Sämdal, Swensson and Wretman 1992; page 391).

Since we are estimating the number of occurrences of rare events, in 50 of the 208 domains, direct estimates of NR and/or of SR are equal to zero, that is,  $\hat{\theta}_{i1} = 0$  and/or  $\hat{\theta}_{i2} = 0$ . Zero point estimates imply that  $\hat{V}(\hat{\theta}_{i1}) = 0$  and/or  $\hat{V}(\hat{\theta}_{i2}) = 0$ , where  $\hat{V}(\hat{\theta}_{i1})$  and  $\hat{V}(\hat{\theta}_{i2})$  are the standard design-based variance estimates of  $\hat{\theta}_{i1}$  and  $\hat{\theta}_{i2}$ , respectively. This result gives a false impression of high accuracy, whereas the exact opposite is more likely to be true in a small area context. Moreover, design based estimates of NR and/or of SR equals to zero produce  $\text{C}\hat{\text{O}}\text{V}(\hat{\theta}_{i1}, \hat{\theta}_{i2}) = 0$ , where  $\text{C}\hat{\text{O}}\text{V}(\hat{\theta}_{i1}, \hat{\theta}_{i2}) = 0$  denotes the standard design-based estimate of the design-based covariance between  $\hat{\theta}_{i1}$  and  $\hat{\theta}_{i2}$ . As a result, covariances also need to be smoothed in a multivariate SAE model.

We hereafter refer to the set of the 50 small areas having one or both zero estimated variances and zero covariances as the "Zero Count" (ZC) set. The complementary set of 158 domains, where  $\hat{V}(\hat{\theta}_{i1}) > 0$  and  $\hat{V}(\hat{\theta}_{i2}) > 0$ , is named the "Non Zero Count" (NZC) set.

Considering the data generating process and the nature of the outcome variables, we expect mainly negative correlations between  $\theta_{i1}$  and  $\theta_{i2}$ . Briefly, we need a suitable distribution for both smoothing covariance matrices and modeling small area parameters that allows for an unrestricted covariance matrix, that is, for both positive and negative correlations.

**Table 1**  
**Variables defining domains of interest**

LLMAAs grouped by productive specialization	Firm size <sup>(b)</sup>	Industrial sector <sup>(a)</sup>
<i>Industrial district</i> <sup>(a,c)</sup>	1-9	1 Food, beverages and tobacco
Food, beverages and tobacco	10-49	2 Textiles and clothing
Textiles and clothing	50-99	3 Paper products, printing and publishing
Paper products, printing and publishing	$\geq 100$	4 Machinery
Machinery		5 Chemicals and basic metals
Jewellery, musical instruments, games, etc.		6 Leather and footwear
Leather and footwear		7 Wood, furniture and household equipment
Wood, furniture and household equipment		8 Jewellery, musical instruments, games, etc.
<i>LLMAAs not defined as district</i> <sup>(c)</sup>		9 Builders, contractors
Non-specialised manufacturing		10 Other manufacturing
Non-specialized, excluding manufacturing		
Tourist		
Cities		

(a) As defined by the 2-digit ATECO 91-ISC 3 level classification and by Sforzi (1991).

(b) Defined according to the number of employees.

(c) Defined in accordance with Istat (1997).



#### 4. An integrated multivariate small area model for count data

Multivariate count data can have a non-trivial correlation structure. In general, the modeling of this structure significantly affects the estimators' efficiency and the computation of correct standard errors. A number of multivariate models for count data have been proposed in the literature, such as the Multivariate Poisson, Multivariate Negative Binomial and Multivariate Poisson-Gamma Mixture models (for a review of such models, see Winkelmann 2003). Unfortunately, these distributions are not suitable for modeling our data since they are based on the hypothesis that correlation is the result of an individual factor that does not vary across outcomes, thus implying a covariance structure restricted to non-negative correlations. In the bivariate case, a more flexible covariance structure is provided by the Latent Poisson Normal distribution (van Ophem 1999); however, any extensions to higher dimensional multivariate data appear impractical.

Aitchison and Ho (1989) proposed a  $d$ -variate distribution that allows for an unrestricted covariance structure, the Multivariate Poisson-Log Normal distribution (MPLN). No closed form exists for this distribution, but it can be represented as a simple mixture allowing for parameter estimation in an MCMC approach (Chib and Winkelmann 2001). Details of the MPLN distribution are provided in the Appendix.

##### 4.1 Smoothing sampling covariance matrices

As previously mentioned, the instability of standard errors in SAE is usually dealt with using a GVF approach. In this section, we present a GVF model with a regression function inspired by the MPLN distribution.

Let  $\mathbf{y}_{il} = [y_{il1}, y_{il2}]'$  be the vector of the two outcome variables referring to the  $l^{\text{th}}$  unit in the  $i^{\text{th}}$  domain. Let  $\mathbf{y}_{il} | \lambda_i, \Sigma_i \perp \mathbf{y}_{il} | \lambda_i, \Sigma_i$  and  $\mathbf{y}_{il} | \lambda_i, \Sigma_i \sim \text{PLN}_2(\lambda_i, \Sigma_i)$ ,  $\forall i, \forall l$ . Under these hypotheses, the moments leading up to the second order can be expressed as follows:

$$E(y_{ijl} | \lambda_i, \Sigma_i) = \exp(\lambda_{ij} + \sigma_{i,jj}/2) = \zeta_{ij}$$

$$V(y_{ijl} | \lambda_i, \Sigma_i) = \zeta_{ij} + \zeta_{ij}^2 [\exp(\sigma_{i,jj}) - 1]$$

$$\text{COV}(y_{ijl}, y_{ihl} | \lambda_i, \Sigma_i) = \zeta_{i1} \zeta_{i2} [\exp(\sigma_{i,jh}) - 1], \quad j \neq h$$

where  $\sigma_{i,jh}$  denotes the  $(j, h)$ ,  $j, h = 1, 2$ , element of  $\Sigma_i$ .

To deal with the problem of smoothing covariance matrices, Otto and Bell (1995), suggested an approach based on a Wishart distributional assumption; specifically, they used smoothed estimates in a small area Normal-Normal model. In the same spirit, we propose a Bayesian approach using the following GVF strategy. Under simple random

sampling, let us assume that the sampling covariance matrix in domain  $i$ ,  $\mathbf{C}_i$  follows a Wishart distribution with  $n_i - 1$  degrees of freedom:

$$\mathbf{C}_i | n_i, \Gamma_i \sim W_2(n_i - 1, \Gamma_i)$$

where  $\Gamma_i = E(\mathbf{C}_i | n_i, \Gamma_i)$ ,  $i = 1, 2, \dots, 158$ , and elements  $(j, h)$  of  $\mathbf{C}_i$  are defined as  $C_{i,jh} = n_i^{-1} \sum_{l=1}^{n_i} (y_{ijl} - \bar{y}_{ij})(y_{ihl} - \bar{y}_{ih})$ , where  $\bar{y}_{ij} = n_i^{-1} \sum_{l=1}^{n_i} y_{ijl}$ .

If  $\zeta_{ij}$  parameters are known, then  $E(\mathbf{C}_i | n_i, \Gamma_i)$  only depends on elements of the  $\Sigma_i$  matrix. We propose to estimate  $\zeta_{ij}$  using the design based estimator  $\hat{\zeta}_{ij} = N_i^{-1} \hat{\theta}_{ij}$ . Thus, we can express each element of the  $\Gamma_i$  matrix as a function of estimates  $\hat{\zeta}_{ij}$  and of the elements of the  $\Sigma_i$  matrix:

$$\Gamma_{i,11} = \hat{\zeta}_{i1} + \hat{\zeta}_{i1}^2 (\exp(\sigma_{i,11}) - 1)$$

$$\Gamma_{i,22} = \hat{\zeta}_{i2} + \hat{\zeta}_{i2}^2 (\exp(\sigma_{i,22}) - 1)$$

$$\Gamma_{i,12} = \hat{\zeta}_{i1} \hat{\zeta}_{i2} (\exp(\sigma_{i,12}) - 1)$$

where  $\sigma_{i,11} = \bar{\sigma}'_{11} \mathbf{Z}_i$ ,  $\sigma_{i,22} = \bar{\sigma}'_{22} \mathbf{Z}_i$ ,  $\sigma_{i,12} = \bar{\sigma}'_{12} \mathbf{Z}_i$ , being  $\mathbf{Z}_i$  is a  $3 \times 1$  vector of dummy variables identifying the firm's size class in the domain  $i$ , and

$$\bar{\sigma}_{11} = \begin{pmatrix} \bar{\sigma}_{1,11} \\ \bar{\sigma}_{2,11} \\ \bar{\sigma}_{3,11} \end{pmatrix}, \bar{\sigma}_{22} = \begin{pmatrix} \bar{\sigma}_{1,22} \\ \bar{\sigma}_{2,22} \\ \bar{\sigma}_{3,22} \end{pmatrix}, \bar{\sigma}_{12} = \begin{pmatrix} \bar{\sigma}_{1,12} \\ \bar{\sigma}_{2,12} \\ \bar{\sigma}_{3,12} \end{pmatrix}$$

that is, we assume that parameters  $\Sigma_i$  are equal for domains belonging to the same firm size class.

We estimate  $\bar{\sigma}_{11}, \bar{\sigma}_{22}, \bar{\sigma}_{12}$  parameters on NZC data. Since we are following a Bayesian approach, prior specifications for  $\bar{\sigma}_{k,jj}$  and  $\bar{\sigma}_{k,12}$   $k = 1, 2, 3$  are needed. We use the following prior specifications:  $\bar{\sigma}_{k,11}^{1/2} \sim U^+$ ,  $\bar{\sigma}_{k,22}^{1/2} \sim U^+$ ,  $\bar{\rho}_k \sim U(-1, 1)$ , where  $\bar{\sigma}_{k,12} = \bar{\rho}_k (\bar{\sigma}_{k,11} \bar{\sigma}_{k,22})^{1/2}$  and  $U^+$  denotes a uniform distribution over a subset of  $R^+$  with a large but finite length. In section 4.3, we show how these estimates can be used to integrate the SAE model with a model for sampling error covariance matrices.

##### 4.2 A Multivariate Normal-Poisson-Log Normal small area model

In this section, we propose a multivariate SAE model based on the MPLN distribution in order to jointly estimate SR and NR using the NZC set.

Let  $\theta_i = (\theta_{i1}, \theta_{i2})'$  be the vector of the two parameters of interest for the  $i^{\text{th}}$  domain in the set of NZC data ( $i = 1, \dots, 158$ ), and let  $\hat{\theta}_i$  be the corresponding vector of direct estimates. The SAE model consists of two separate models. The first model is a sampling model:

$$\hat{\theta}_i | \theta_i \sim \text{ind } N_2(\theta_i | \Psi_i), \quad i = 1, \dots, 158. \quad (1)$$



As in Lahiri and Rao (1995), we justify the normality assumption in (1) using the central limit argument. It is standard practice to assume that sampling error covariance matrices  $\Psi_i$  are known, and a GVF method is generally used to estimate  $\Psi_i$ . Here, as a smoothed estimation of  $\Psi_i$ , we adopt  $\hat{\Psi}_i = E(\Gamma_i | C_i, n_i) K_i$ , where  $K_i = N_i (N_{t3i}/n_{t3i} - 1)$ . From this point on we will refer to  $\hat{\Psi}_i$  as Smoothed Sampling Error Covariance matrix (SMSEC).

The second component of the SAE model is a linking model that relates  $\theta_i$  to area specific auxiliary data:

$$\theta_i \sim \text{ind PLN}_2(\eta_i, \Sigma_v), \quad i = 1, \dots, 158,$$

where (2)

$$\eta_i = \alpha + \gamma Z_i + \beta Z_i x_i$$

$Z_i$  is a  $3 \times 1$  vector of dummy variables identifying the firm's size class in the domain  $i$  and  $x_i = \log(x_i^*)$ , where  $x_i^*$  is the number of employees in the domain  $i$ .

At the end,  $\Sigma_v$  is the covariance matrix related to the area-specific random effects:

$$\Sigma_v = \begin{pmatrix} \sigma_{v,11} & \sigma_{v,12} \\ \sigma_{v,21} & \sigma_{v,22} \end{pmatrix}$$

and

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \gamma = \begin{pmatrix} 0 & \gamma_{12} & \gamma_{13} \\ 0 & \gamma_{22} & \gamma_{23} \end{pmatrix}, \beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix}.$$

From here on, we refer to this small area model as "Multivariate Normal-Poisson-Log Normal" (MNPLN).

We adopt a fully hierarchical Bayesian approach. In this framework, relatively complex (e.g., multivariate) models can be implemented easily; in addition, posterior distributions can be approximated using MCMC algorithms. Computing small area multivariate estimates, and estimates of their MSE in particular, can be difficult within a frequentist approach. The specification of priors for the described model is as follows:

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \sim N_2(\mathbf{0}, \sigma^2 \mathbf{I}_2),$$

$$\begin{pmatrix} \gamma_{1k'} \\ \gamma_{2k'} \end{pmatrix} \sim N_2(\mathbf{0}, g_k \mathbf{I}_2) \quad k' = 2, 3,$$

$$\begin{pmatrix} \beta_{1k} \\ \beta_{2k} \end{pmatrix} \sim N_2(\mathbf{0}, b_k \mathbf{I}_2) \quad k = 1, 2, 3,$$

$$\Sigma_v^{-1} \sim W(s, \mathbf{I}_2),$$

$$\begin{pmatrix} \gamma_{1k'} \\ \gamma_{2k'} \end{pmatrix} \perp \begin{pmatrix} \beta_{1k} \\ \beta_{2k} \end{pmatrix},$$

where  $s = 3$  and  $a, g_k, b_k$  are large compared with the scale of the data. This is to reflect the lack of prior information about model parameters, thus defining diffuse but proper specification of priors. The posterior means  $\hat{\theta}_i^{\text{HB}} = E(\theta_i | \hat{\theta}_i, \hat{\Psi}_i)$  are taken as estimators of the area parameters, while the posterior variance  $V(\theta_i | \hat{\theta}_i, \hat{\Psi}_i)$  is used as a measure of uncertainty.

For the sake of comparison, we take the standard Multivariate Normal-Normal (MNN) model as a benchmark, where the sampling model is defined as in (1) and the linking model is defined as follows:

$$\theta_i \sim \text{ind } N_2(\mu_i^*, \Sigma_v^*), \quad (3)$$

where  $\mu_i^* = \alpha^* + \gamma^* Z_i + \beta^* Z_i x_i$ . Parameters  $\alpha^*, \gamma^*, \beta^*$  and their prior distributions are defined as  $\alpha, \gamma$  and  $\beta$  in the previous model.

#### 4.3 An integrated MNPLN small area model

In order to account for the extra variability due to the estimated covariance matrices of sampling errors, as well as to overcome the zero variances and covariances problem, we suggest a solution in the spirit of that proposed by Arora and Lahiri (1997), Liu *et al.* (2007) and You (2008). We integrate the model for sampling error covariance matrices of section 4.1 into SAE models (1) and (2). Thus, we here refer to the whole set of 208 domains.

In this context, the small area sampling model is formulated as usual, that is,  $\hat{\theta}_i | \theta_i \sim \text{ind } N_2(\theta_i, \Psi_i^*)$ ,  $i = 1, \dots, 208$ . Under the hypotheses regarding  $y_{ij}$  formulated in section 4.1, assuming that the  $\Sigma_i$ s are known and assuming that  $\theta_{ij} = N_i \zeta_{ij}$ , the elements of the sampling error covariance matrix  $\Psi_i^*$  can be expressed as follows:

$$\Psi_{i,jj}^* = K_i [\theta_{ij}/N_i + \theta_{ij}^2/N_i^2 (\exp(\hat{\sigma}_{jj}' Z_i) - 1)] \quad (4)$$

$$\Psi_{i,12}^* = K_i [N_i^{-2} \theta_{i1} \theta_{i2} (\exp(\hat{\sigma}_{12}' Z_i) - 1)] \quad (5)$$

where  $\hat{\sigma}_{jj}'$   $j = 1, 2$  and  $\hat{\sigma}_{12}'$  are posterior means of parameters  $\bar{\sigma}_{jj}$  and  $\bar{\sigma}_{12}$ , respectively, computed using the model of section 4.1.

Since the sampling error covariance matrices are expressed as a function of the  $\theta_i$  parameters, here they can be considered Model Based Sampling Error Covariances (MBSEC). The posterior means  $\hat{\theta}_i^{\text{HB}} = E(\theta_i | \hat{\theta}_i)$  are taken as estimators of  $\theta_i$ s, while the posterior variance  $V(\theta_i | \hat{\theta}_i)$  is used as a measure of uncertainty.

We note that the MNN model cannot be implemented following the integrated approach described above. In fact, (3) does not ensure the positivity of  $\theta_i$  nor of the diagonal elements of  $\Psi_i$  as a result.

## 5. Data analysis

In section 5.1, we compare the MNPLN model with the benchmark MNN model and their univariate counterparts. We assume SMSEC for both models; we thus refer to the two strategies as MNPLN-SMSEC and MNN-SMSEC from here on. Since these models do not allow us to deal with the zero count problem, we refer this analysis to the NZC set. In section 5.2, we compare the SAE integrated strategy based on the MNPLN model and MBSEC (MNPLN-MBSEC), which we presented in Section 4.3, with the strategy based on the MNPLN-SMSEC. We limit the analysis to the NZC set in order to evaluate the two strategies under the same conditions. Finally, in section 5.3 we evaluate the overall performance of the proposed SAE model MNPLN-MBSEC for the whole data set (NZC+ZC).

Posterior distributions of parameters were obtained for all models, using Monte Carlo integration via the Gibbs sampling algorithm. We used the MCMC software WinBUGS (Spiegelhalter, Thomas, Best and Gilks 1995) to run three parallel chains (each with 25,000 runs), the starting point being drawn from an over-dispersed distribution. WinBUGS codes are available at the URL <http://www2.stat.unibo.it/trivisano/>. The convergence of the Gibbs sampler was monitored by visual inspection of the chains' plots and of autocorrelation diagrams, and by means of the potential scale reduction factor proposed by Gelman and Rubin (1992). Although all models displayed fast convergence, we discarded the first 5,000 iterations from each chain. In multivariate models, the fairly strong autocorrelation of chains is reduced by thinning the chain (1 out of every 3 values has been considered for posterior summaries). See Rao (2003, pages 228-232) for details.

The performances of the small area models discussed in sections 4.2 and 4.3 are compared using various measures. In order to choose among competing models, we computed the Deviance Information Criterion (DIC). The DIC is a model selection criterion according to which a model's performance is evaluated as the sum of a measure of fit (the posterior mean of the deviance  $\bar{D}$ ) and a measure of complexity obtained as the difference between  $\bar{D}$  and the deviance evaluated at the parameters' posterior mean. In this way, a model is preferred if it displays a lower DIC value (Spiegelhalter, Best, Carlin and Van der Linde 2002).

In order to verify the strength of the multivariate approach to SAE, we use as a benchmark the univariate versions of models discussed in sections 4.2 and 4.3, defined as follows. For all models, we set  $\sigma_{v,12} = 0$  in  $\Sigma_v$ , and we assume  $\sigma_{v,11} \perp \sigma_{v,22}$ ,  $\sigma_{v,jj}^{1/2} \sim U(0, U^+)$ ,  $j = 1, 2$ . For SMSEC models, we set  $\Psi_j = \text{diag}(\hat{\Psi}_j)$ , while for MBSEC models we set  $\sigma_{1,12} = 0$  in (5). In addition, a new

set of estimates for parameters  $\bar{\sigma}_{11}$  and  $\bar{\sigma}_{22}$  is obtained by setting  $\bar{\rho}_k = 0$  in the model of section 4.1.

Table 2 reports the DIC results for the whole set of small area models.

**Table 2**  
Model comparison using DIC statistic

Model	Data set	DIC
MNN-SMSEC (univariate version)	NZC	2,742.2
	NZC	2,745.4
MNPLN-SMSEC (univariate version)	NZC	2,656.9
	NZC	2,661.0
MNPLN-MBSEC (univariate version)	NZC	2,623.6
	NZC	2,638.1
MNPLN-MBSEC (univariate version)	NZC+ZC	3,202.7
	NZC+ZC	3,214.3

All the multivariate models considered perform better in terms of DIC than their univariate counterparts (Table 2). In addition, for all multivariate models we find that posterior credibility intervals of  $\rho_v = \sigma_{v,12}/\sqrt{\sigma_{v,11}\sigma_{v,22}}$  do not contain zero. We thus focus on multivariate models in the following paragraphs.

We checked the adequacy of the specified multivariate models using posterior predictive checks. Simulated values of a suitable discrepancy measure are generated from the posterior predictive distribution and are then compared with the values of the same measure computed from observed data. Let  $\hat{\theta}_{\text{obs}}$  and  $\hat{\theta}_{\text{new}}$  denote the observed and generated data, respectively. The posterior predictive  $p$ -value is defined as  $p = P\{d(\hat{\theta}_{\text{new}}, \theta) > d(\hat{\theta}_{\text{obs}}, \theta) \mid \hat{\theta}_{\text{obs}}\}$ . We consider a discrepancy measure proposed in Datta *et al.* (1999), which is defined as

$$d(\hat{\theta}, \theta) = \sum_{i=1}^N (\hat{\theta}_i - \theta_i)' \Psi^{-1} (\hat{\theta}_i - \theta_i). \quad (6)$$

Computing the  $p$ -value is straightforward using the MCMC output. Extreme values of the probability  $p$  indicate a given model's lack of fit. Following Rao (2003, page 245-246) and You and Rao (2002), we computed two statistics that are useful in order to assess model fit at the individual domain level. The first statistic,  $p_{ij}^* = P(\hat{\theta}_{ij, \text{new}} < \hat{\theta}_{ij, \text{obs}} \mid \hat{\theta}_{\text{obs}})$ , provides information about the degree of consistent overestimation or underestimation of  $\hat{\theta}_{ij, \text{obs}}$ .

The second statistics is defined as

$$d_{ij}^* = [E(\hat{\theta}_{ij} \mid \hat{\theta}_{\text{obs}}) - \hat{\theta}_{ij, \text{obs}}] / \sqrt{V(\hat{\theta}_{ij} \mid \hat{\theta}_{\text{obs}})},$$

where expectation and variance are under the posterior predictive distribution. Table 3 summarizes results relative to  $p$ ,  $p_{ij}^*$  and  $d_{ij}^*$ .

To further check the consistency of the data, we calculated direct and model-based estimates of  ${}_A\theta_{sj}$ ,  $s = 1, \dots, 10$ , that is, the total number of NR and SR for the ten domains identified by classifying firms only according to the industrial sector. Let  $w_{is} = 1$  if the number of recruits in the domain  $i$  refers to the industrial sector  $s$  and  $w_{is} = 0$ ; otherwise, then

$${}_A\theta_{sj} = \sum_i \theta_{ij} w_{is} \quad (7)$$

At this level of aggregation, direct estimates can be considered accurate. Consequently, given two sets of model-based estimates referring to these large domains, we prefer the one that agrees with the direct estimates. Domains identified by industrial sectors are planned in the Excelsior Survey; each industrial sector is stratified according to firm size. Therefore, direct estimates  ${}_A\hat{\theta}_{sj}$  for each industrial sector are calculated using the standard Horwitz-Thompson estimator. Aggregated model-based estimates are computed based on the MCMC output. For models referring to NZC data, we aggregated following (7) at each MCMC step  $t$ ,  $t = 1, \dots, T$ , with samples  $'\theta_{ij}^*$  and  $'\theta_{ij}^{**}$  generated respectively from the posterior distribution of  $\theta_{ij}$  for domains belonging to the NZC set and from the predictive distribution of  $\theta_{ij}$  for domains belonging to the ZC set. The HB estimator is defined as  ${}_A\hat{\theta}_{sj}^{\text{HB}} = T^{-1} \sum_{t=1}^T (\sum_{i \in \text{NZC}} '\theta_{ij}^* w_{is} + \sum_{i \in \text{ZC}} '\theta_{ij}^{**} w_{is})$ . Otherwise, for the model on NZC+ZC data, we aggregated following (7) MCMC samples from the posterior distributions of  $\theta_{ij}$ . In this case, the HB estimator is defined as  ${}_A\hat{\theta}_{sj}^{\text{HB}} = T^{-1} \sum_{t=1}^T (\sum_{i \in \text{NZC}} '\theta_{ij}^* w_{is})$ . Table 4 reports summaries of  ${}_A\hat{\theta}_{sj}$  and  ${}_A\hat{\theta}_{sj}^{\text{HB}}$ .

For all the multivariate models, we examined the following variants of the prior distributions: independent non-informative flat prior distributions were used for the elements of vectors  $\alpha, \beta, \gamma, \alpha^*, \beta^*$ , and  $\gamma^*$ ;  $\sigma_{v,jj}^{1/2} \sim U^+$ ,  $j = 1, 2$ ,  $\rho_v \sim U(-1, 1)$ ,  $\sigma_{v,12} = \rho_v (\sigma_{v,12} \sigma_{v,12})^{1/2}$ . We do the same for the elements of matrix  $\Sigma^*$  in the MNN model. We did not find any relevant changes in the posterior distributions of parameters of interest.

### 5.1 Comparing the MNPLN-SMSEC and MNN-SMSEC models on the NZC set

We find that the MNPLN-SMSEC model largely outperforms the MNN-SMSEC one in terms of DIC (Table 2). This last model shows a lack of fit as it displays a  $p$ -value equal to 0.034 (Table 3), whereas a value of 0.65 suggests the adequacy of the MNPLN-SMSEC model. This finding is confirmed when  $p_{ij}^*$  and  $d_{ij}^*$  measures (Table 3) for the two models are compared. For the MNN-SMSEC model,  $p_{ij}^*$  ranges over domains from 0.000 to 0.995 for NR ( $j = 1$ ) and from 0.003 to 0.993 for SR ( $j = 2$ ), respectively, indicating overestimation and underestimation in some domains. In addition, summaries of the standardized residuals  $d_{ij}^*$  indicate that there are predicted values outside two standard deviations of the corresponding observed values. The same measures for the MNPLN-SMSEC model indicate an adequate fit.

We also find that the MNPLN-SMSEC model outperforms the MNN-SMSEC models when performances are evaluated with reference to estimates for large domains (Table 4). In fact, credibility intervals for the MNN-SMSEC only cover 2 aggregated direct estimates for NR and 4 for SR, while credibility intervals under the MNPLN-SMSEC cover 6 aggregated direct estimates for NR and 6 for SR.

**Table 3**  
Posterior predictive checks; summaries of  $p_{ij}^*$  and  $d_{ij}^*$  calculated with respect to  $i$

Model	Data set	$p$		$p_{i1}^*$	$p_{i2}^*$	$d_{i1}^*$	$d_{i2}^*$
MNN-SMSEC	NZC	0.034	min	0.000	0.003	-3.764	-2.867
			median	0.591	0.616	0.257	0.295
			max	0.995	0.993	2.656	-2.515
MNPLN-SMSEC	NZC	0.65	min	0.154	0.129	-0.965	-1.165
			median	0.535	0.561	0.124	0.149
			max	0.891	0.912	1.216	1.286
MNPLN-MBSEC	NZC	0.78	min	0.090	0.134	-1.085	-0.983
			median	0.515	0.519	-0.084	-0.085
			max	0.916	0.914	1.401	1.787
MNPLN-MBSEC	NZC+ZC	0.79	min	0.072	0.111	-1.164	-0.945
			median	0.506	0.523	-0.076	-0.094
			max	0.903	0.913	1.301	1.778



**Table 4**  
Direct and HB estimates for industrial sectors; in italic HB estimates whose credibility intervals cover direct estimates

Direct estimates			HB estimates											
s			MNN-SMSEC (NZC)			MNPLN-SMSEC (NZC)			MNPLN-MBSEC (NZC)			MNPLN-MBSEC (NZC+ZC)		
	$\hat{\theta}_{s1}$	$se(\hat{\theta}_{s1})$	$\hat{\theta}_{s1}^{HB}$	95% cred int.		$\hat{\theta}_{s1}^{HB}$	95% cred int.		$\hat{\theta}_{s1}^{HB}$	95% cred int.		$\hat{\theta}_{s1}^{HB}$	95% cred int.	
1	1,702.0	41.3	1,077.0	964.3	1,201.0	1,266.0	1,055.0	1,509.0	1,649.0	1,434.0	1,906.0	1,630.0	1,406.0	1,899.0
2	1,758.8	41.9	1,936.0	1,793.0	2,091.0	1,720.0	1,441.0	2,011.0	1,975.0	1,665.0	2,347.0	1,908.0	1,598.0	2,291.0
3	725.0	26.9	557.8	460.6	662.7	534.6	435.8	642.3	696.6	573.3	842.3	682.8	575.5	811.8
4	373.9	19.3	202.7	123.0	294.8	192.1	129.1	277.0	370.0	291.1	471.4	319.8	252.1	408.3
5	142.4	11.9	158.2	66.5	258.2	146.0	98.4	205.7	235.6	164.3	326.9	149.7	108.3	205.0
6	5,624.1	75.0	4,134.0	3,800.0	4,484.0	5,235.0	4,814.0	5,670.0	5,537.0	5,136.0	5,963.0	5,594.0	5,187.0	6,029.0
7	887.7	29.8	659.9	549.1	783.7	629.6	526.4	743.4	872.7	761.7	1,003.0	844.6	732.3	980.3
8	223.9	15.0	263.3	188.2	340.6	260.6	182.8	351.3	362.0	262.8	494.1	288.7	203.1	410.8
9	661.5	25.7	893.7	790.3	999.4	777.6	624.7	948.7	931.0	754.8	1,150.0	803.3	638.7	1,017.0
10	1,792.6	42.3	1,460.0	1,334.0	1,598.0	1,579.0	1,381.0	1,798.0	1,847.0	1,650.0	2,074.0	1,813.0	1,610.0	2,053.0
	$\hat{\theta}_{1,2}$	$se(\hat{\theta}_{1,2})$	$\hat{\theta}_{1,2}^{HB}$	95% cred int.		$\hat{\theta}_{1,2}^{HB}$	95% cred int.		$\hat{\theta}_{1,2}^{HB}$	95% cred int.		$\hat{\theta}_{1,2}^{HB}$	95% cred int.	
1	942.7	300.2	482.0	428.5	531.3	503.7	413.3	600.4	832.6	706.4	987.6	817.8	686.0	980.0
2	920.0	135.7	883.9	798.7	967.4	849.8	694.8	1,022.0	949.8	778.9	1,161.0	922.3	747.6	1,167.0
3	253.2	35.6	249.2	209.2	292.1	254.1	202.1	309.9	338.8	269.2	423.1	284.7	226.2	354.5
4	150.5	36.0	84.4	53.3	120.4	84.7	56.8	119.2	160.6	116.7	218.0	131.5	97.0	179.6
5	39.8	16.6	66.7	31.2	104.2	62.0	37.3	89.3	116.3	74.3	173.0	60.9	38.4	90.5
6	2,304.0	131.5	1,869.0	1,692.0	2,054.0	2,070.0	1,856.0	2,282.0	2,273.0	2,060.0	2,508.0	2,297.0	2,079.0	2,542.0
7	532.7	105.8	293.0	247.7	345.6	299.0	245.9	357.2	471.5	402.8	553.2	443.3	377.2	538.3
8	80.8	32.3	115.7	85.7	143.5	100.5	67.7	140.3	139.5	76.7	210.4	98.0	58.5	156.9
9	362.7	66.3	407.0	358.6	453.0	361.0	285.8	438.8	432.1	335.4	552.9	360.4	274.7	476.2
10	856.3	70.7	661.1	598.1	722.6	714.4	614.0	824.7	855.4	740.5	984.6	832.7	719.8	964.5

## 5.2 Comparing the MNPLN-SMSEC and MNPLN-MBSEC models on the NZC set

Values of  $p$ ,  $p_{ij}^*$  and  $d_{ij}^*$  are approximately comparable for the MNPLN-SMSEC and MNPLN-MBSEC models (Table 3). Likewise, model-based estimates produced by MNPLN-SMSEC assume values very close to those obtained using MNPLN-MBSEC; in fact, the correlation between the posterior means of  $\theta_{i1}$  under the two models is equal to 0.98, while the same measure referring to  $\theta_{i2}$  is equal to 0.94. The same results arise for the correlation between posterior standard errors, which are 0.92 and 0.94, respectively. Performances of the MNPLN-MBSEC model in terms of agreement with direct estimates of large domains (Table 4) are slightly better than those of the MNPLN-SMSEC model: respectively, 7 direct estimates of NR and 8 of SR are covered by the credibility interval calculated under this model.

Given these results, we conclude that the fit of the MNPLN-MBSEC model is adequate.

## 5.3 Evaluating the performances of MNPLN-MBSEC models on the NZC+ZC set

We observe that the performances of the MNPLN-MBSEC model on the whole dataset in terms of  $p$ ,  $p_{ij}^*$  and  $d_{ij}^*$  measures are satisfactory and comparable with those of the same model on the NZC data set (Table 3). Obviously, DIC values for the two models cannot be compared as the two models are estimated on different data sets.

As can be seen in Table 4, all the credibility intervals calculated using this model cover direct estimates referring to large domains; in other words, the agreement of HB estimates with direct estimates is very satisfactory. This result can be explained by noting that zero counts are more probable in small domains, which are characterized by a small number of employees (the covariate in all models). Therefore, estimating models on NZC data can lead to biased estimates of parameter  $\beta$ . We conclude that integrating a sampling covariance model into the MNPLN small area model leads to an appreciable increase in the reliability

of small area estimates. To describe the efficiency gain of the HB estimates, we computed on the NZC set the average percent CV reduction (You 2008), defined as the average of the difference of the direct CV and HB CV (the ratio of the square root of the posterior variance and the posterior mean) relative to direct CV. The average CV reduction is 23.1% for NR and 29.1% for SR.

### Acknowledgements

The authors would like to thank the Editor, Associate Editor and Referee for their helpful comments and suggestions. The research leading to this paper was partially supported by Miur-PRIN 2003/2003133249 and Miur-Prin 2008/2008CEFF37-001.

### Appendix

#### The Multivariate Poisson-Log Normal distribution

Let  $\mathbf{y} = (y_1, y_2, \dots, y_j, \dots, y_d)$  be a  $d$ -dimensional vector of counts, and suppose that  $y_j | \tau_j \sim \text{Po}(\tau_j)$ , with  $y_j | \tau_j \perp y_{j'} | \tau_{j'} (j \neq j')$ . Let the vector of parameters  $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_j, \dots, \tau_d)$  follow a multivariate Log Normal, that is,  $\boldsymbol{\tau} | \boldsymbol{\lambda}, \boldsymbol{\Sigma} \sim \text{LN}_d(\boldsymbol{\lambda}, \boldsymbol{\Sigma})$ , where  $\boldsymbol{\lambda} = E(\log \boldsymbol{\tau})$  and  $\boldsymbol{\Sigma} = \text{COV}(\log \boldsymbol{\tau})$ . Then the marginal distribution of  $\mathbf{y}$  is a Multivariate Poisson-Log Normal (MPLN) distribution, which is a log normal mixture of  $d$  independent  $\text{Po}(\tau_j)$ , that is,  $\mathbf{y} | \boldsymbol{\lambda}, \boldsymbol{\Sigma} \sim \text{PLN}_d(\boldsymbol{\lambda}, \boldsymbol{\Sigma})$ . By denoting the  $(j, h)$ ,  $j, h = 1, 2, \dots, d$  element of  $\boldsymbol{\Sigma}$  as  $\sigma_{jh}$ , marginal moments can be obtained easily through conditional expectation results and the standard properties of the Poisson and Log Normal distributions:

$$E(y_j | \boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \exp(\lambda_j + \sigma_{jj}/2) = \zeta_j$$

$$V(y_j | \boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \zeta_j + \zeta_j^2 [\exp(\sigma_{jj}) - 1]$$

$$\text{COV}(y_j, y_h | \boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \zeta_j \zeta_h [\exp(\sigma_{jh}) - 1], \quad j \neq h.$$

Note that the MPLN model allows for overdispersion provided that  $\sigma_{jj} > 0$ , thus leading to  $V(y_j | \boldsymbol{\lambda}, \boldsymbol{\Sigma}) > E(y_j | \boldsymbol{\lambda}, \boldsymbol{\Sigma})$ . Moreover, the correlation structure of counts is unrestricted, since  $\text{COV}(y_j, y_h | \boldsymbol{\lambda}, \boldsymbol{\Sigma})$  can be either positive or negative depending on the sign of  $\sigma_{jh}$ . Aitchison and Ho (1989), as well as Good and Pirog-Good (1989), studied a bivariate MPLN distribution, albeit exclusively in cases without covariates. However, the same model can easily be extended to take covariates into consideration (Chib and Winkelmann 2001).

### References

- Aitchison, J., and Ho, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76, 643-653.
- Arora, V., and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.
- Baldi, C., Bellisai, D., Fivizzani, S. and Sorrentino, M. (2007). Production of job vacancy statistics: Coverage. *Contributi Istat, Istituto Nazionale di Statistica*.
- Becattini, G. (1992). The Marshallian industrial district as a socio-economic notion. In *Industrial Districts and International Co-operation in Italy*, (Eds., F. Pyke, G. Becattini and W. Sengenberger). International Labor Office, Geneva.
- Chattopadhyay, M., Lahiri, P., Larsen, M. and Reimnitz, J. (1999). Composite estimation of drug prevalence for sub-state areas. *Survey Methodology*, 25, 81-86.
- Chen, S. (2001). Empirical best prediction and hierarchical Bayes methods in small area estimation. Ph.D. Dissertation, Department of Mathematics and Statistics, University of Nebraska, Lincoln.
- Chib, S., and Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19, 428-435.
- Cohen, M.L. (2000). Evaluation of Census Bureau's small-area poverty estimates. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 62-68.
- Datta, G.S., Fay, R.E. and Ghosh, M. (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation. *Proceedings of Bureau of the Census 1991 Annual Research Conference*, U. S. Bureau of the Census, Washington, DC, 63-79.
- Datta, G.S., Ghosh, M., Nangia, N. and Natarajan, K. (1996). Estimation of median income of four-person families: A Bayesian approach. In *Bayesian Analysis in Statistics and Econometrics*, (Eds., D.A. Berry, K.M. Chaloner and J.M. Geweke). New York: John Wiley & Sons, Inc., 129-140.
- Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 488, 1074-1082.
- Elazar, D. (2004). Small area estimation of disability in Australia. *Statistics in Transition*, 6, 5, 667-684.
- Fabrizi, E., Ferrante, M.R. and Pacei, S. (2005). Estimation of poverty indicators at sub-national level using multivariate small area models. *Statistics in Transition*, 7, 3, 587-608.
- Fabrizi, E., Ferrante, M.R. and Pacei, S. (2008). Measuring sub-national income poverty by using a small area multivariate approach. *Review of Income and Wealth*, 54, 4, 597-615.
- Fay, R.E. (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics*, (Eds., R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: John Wiley & Sons, Inc., 91-102.

- Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-511.
- Ghosh, M., Nangia, N. and Kim, D. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.
- Good, D.H., and Pirog-Good, M.A. (1989). Models for bivariate count data with an application to teenage delinquency and paternity. *Sociological Methods and Research*, 17, 4, 409-431.
- Istat (1997). I sistemi locali del lavoro 1991. *Argomenti*, Roma 1997, 10.
- Lahiri, P., and Rao, J.N.K. (1995). Robust estimation of mean square error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.
- Liu, B., Lahiri, P. and Kalton, G. (2007). Hierarchical Bayes modeling of survey weighted small area proportions. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3181-3186.
- Otto, M.C., and Bell, W.R. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the Section on Government Statistics*, American Statistical Association, 160-165.
- Rao, J.N.K. (2003). *Small Area Estimation*. New Jersey: John Wiley & Sons, Inc.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model-Assisted Survey Sampling*. New York: Springer-Verlag.
- Sforzi, F. (1991). I distretti industriali marshalliani nell'economia italiana. In *Distretti industriali e cooperazione fra imprese in Italia*, (Eds., F. Pyke, G. Becattini and W. Sengenberger). Quaderni di Studi e Informazioni, 34.
- Sforzi, F., and Lorenzini, F. (2002). I distretti industriali. In *Ministero delle Attività Produttive-IPI, L'esperienza italiana dei distretti industriali*, Roma, IPI.
- Spiegelhalter, D.J., Best, N., Carlin, B.P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64, 583-639.
- Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling*. Version 0.50, Medical Research Council Biostatistics Unit, Cambridge.
- Van Ophem, H. (1999). A general method to estimate correlated discrete random variables. *Econometric Theory*, 15, 228-237.
- Winkelmann, R. (2003). *Econometric Analysis of Count Data*. Springer, Berlin.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- You, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, 1, 19-27.
- You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.
- You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, 30, 3-15.
- You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 25-32.



# Linearization variance estimation for generalized raking estimators in the presence of nonresponse

Julia D'Arrigo and Chris Skinner<sup>1</sup>

## Abstract

Alternative forms of linearization variance estimators for generalized raking estimators are defined via different choices of the weights applied (a) to residuals and (b) to the estimated regression coefficients used in calculating the residuals. Some theory is presented for three forms of generalized raking estimator, the classical raking ratio estimator, the 'maximum likelihood' raking estimator and the generalized regression estimator, and for associated linearization variance estimators. A simulation study is undertaken, based upon a labour force survey and an income and expenditure survey. Properties of the estimators are assessed with respect to both sampling and nonresponse. The study displays little difference between the properties of the alternative raking estimators for a given sampling scheme and nonresponse model. Amongst the variance estimators, the approach which weights residuals by the design weight can be severely biased in the presence of nonresponse. The approach which weights residuals by the calibrated weight tends to display much less bias. Varying the choice of the weights used to construct the regression coefficients has little impact.

Key Words: Calibration; Nonresponse; Raking; Variance estimation; Weight.

## 1. Introduction

Survey weighting is widely used to adjust for non-response bias. Generalized raking estimation (Deville, Särndal and Sautory 1993) provides a class of weighting methods which may be used when population totals of auxiliary variables are available. These methods can, in principle, remove (large-sample) nonresponse bias when the probability of nonresponse is related to the values of the auxiliary variables via a generalized linear model.

This paper presents some theory for linearization variance estimation for such methods in the presence of nonresponse. It also reports a simulation study of the properties of alternative raking estimators and associated variance estimators in settings designed to mimic two European surveys conducted by national statistical institutes. We consider three forms of raking estimator: the classical raking ratio estimator, the 'maximum likelihood' raking estimator (Brackstone and Rao 1979; Fuller 2002) and the generalized regression estimator (GREG). The first estimator has been used in practice in the British Labour Force Survey (LFS), the first survey upon which our simulation study is based. A version of the second estimator has been used in practice in the German Survey of Income and Expenditure (SIE), the second survey upon which our simulation study is based. The GREG estimator is widely used in many surveys, in particular in the context of nonresponse (Särndal and Lundström 2005).

A number of weighting methods, which do not fall into the class of generalized raking methods considered here, have also been proposed. See Särndal and Lundström (2005) for a historical account and Kott (2006) and Chang and Kott (2008) for some recent developments where the

auxiliary variables for which population-level information is available may differ from those variables which are used as covariates in the generalized linear model for the probability of nonresponse.

The primary focus of this paper is on variance estimation and specifically on linearization methods, for which there exist a number of slightly different forms of variance estimator in the literature. In our simulation study we shall compare the properties of alternative raking estimators and associated variance estimators with respect to the effects of both sampling and nonresponse. A previous simulation study by Stukel, Hidiroglou and Särndal (1996) found little difference between two forms of linearization estimator with respect to sampling. However, there are reasons why non-response may lead to greater differences. Conditions for unbiasedness of raking estimation methods under non-response models vary between estimation methods (*e.g.*, Kalton and Maligalig 1991; Kalton and Flores-Cervantes 2003) and the choice of variance estimator may be more important in the presence of nonresponse (*e.g.*, Fuller 2002, Section 8).

The paper is structured as follows. The generalized raking estimators are defined in section 2 and, after introducing an asymptotic framework, the bias of these estimators is considered in section 3. Linearization variance estimators are defined in section 4. The simulation study is presented in section 5, the results are discussed in section 6 and some concluding remarks are given in section 7.

## 2. Generalized raking estimation

We consider the class of weighted estimators of a population total  $T_y = \sum_U y_i$ , which may be expressed as

1. Julia D'Arrigo and Chris Skinner, University of Southampton. E-mail: C.J.Skinner@soton.ac.uk.

$\hat{T}_y = \sum_s w_i y_i$ , where  $y_i$  is the value of a survey variable for a unit  $i$  in a sample  $s$  from a population  $U$  and  $w_i$  is the *survey weight* which may depend on the sample but not on the choice of survey variable. We suppose here that the sample  $s$  consists of the set of respondents remaining after sampling and possible unit nonresponse. Generalized raking is a form of weighted estimation which may be employed when auxiliary population information is available in the form of a vector  $T_x = \sum_U x_i$  of population totals of values  $x_i$  of a vector of auxiliary variables, where  $x_i$  is known for all units in  $s$ . Following Deville and Särndal (1992), the weights  $w_i$  are said to be *calibrated* if they satisfy the *calibration equations*  $\sum_s w_i x_i = T_x$ . The vector  $T_x$  is referred to as the vector of *calibration totals*. The class of generalized raking weights  $w_i$  is obtained by minimising the objective function:

$$\sum_i d_i G(w_i / d_i), \quad (2.1)$$

subject to the weights  $w_i$  being calibrated, where  $G(\cdot)$  is a specified objective function which meets certain criteria (see Deville *et al.* 1993) and  $d_i$  is an initial weight. We shall take this to be the design weight, *i.e.*,  $d_i = \pi_i^{-1}$ , where  $\pi_i$  is the probability that unit  $i$  is sampled. Deville and Särndal (1992) show that (subject to  $G(\cdot)$  obeying certain conditions), the solution of the above constrained optimisation problem may be expressed as:

$$w_i = d_i F(x_i' \hat{\lambda}), \quad (2.2)$$

where  $F(u) = g^{-1}(u)$  denotes the inverse function of  $g(u) = dG(u)/du$  and  $\hat{\lambda}$  is the Lagrange multiplier which solves the calibration equations:

$$\sum_s d_i F(x_i' \hat{\lambda}) x_i = T_x. \quad (2.3)$$

Deville and Särndal (1992) discuss various choices of the  $G(\cdot)$  function and associated  $F(\cdot)$  function. We consider the following three choices:

*linear:*

$$G_L(u) = (1/2)(u-1)^2, \quad F_L(u) = 1+u;$$

*multiplicative (raking ratio):*

$$G_M(u) = u \log(u) - u + 1, \quad F_M(u) = \exp(u);$$

*maximum likelihood raking:*

$$G_{ML}(u) = u - 1 - \log(u), \quad F_{ML}(u) = (1-u)^{-1}.$$

See also Deville *et al.* (1993) and Fuller (2009, section 2.9) regarding the above terminology for these functions. With the linear choice of  $G(\cdot)$ , the optimisation problem has a closed form solution and the generalized raking estimator

becomes  $\hat{T}_y = \hat{T}_{yd} + (T_x - \hat{T}_{xd})' \hat{B}_s$ , the *generalised regression estimator* (GREG), where  $\hat{T}_{yd} = \sum_s d_i y_i$ ,  $\hat{T}_{xd} = \sum_s d_i x_i$  and

$$\hat{B}_s = \left( \sum_s d_i x_i x_i' \right)^{-1} \sum_s d_i x_i y_i. \quad (2.4)$$

With the multiplicative choice of  $G(\cdot)$ , the calibrated estimator of  $T_y$  is the classical raking ratio estimator (Brackstone and Rao 1979) when  $T_x$  contains the population counts in the categories of two or more categorical auxiliary variables. For example, in the context of the Britain Labour Force Survey,  $x_i$  denotes the vector of indicator variables of three categorical auxiliary variables:  $x_i = (\delta_{1..i}, \dots, \delta_{A..i}, \delta_{1.i}, \dots, \delta_{B.i}, \delta_{..1i}, \dots, \delta_{..Ci})'$ , where  $\delta_{a..i} = 1$  if unit  $i$  is in category  $a$  of the first auxiliary variable and 0 otherwise,  $\delta_{.b.i} = 1$  if unit  $i$  is in category  $b$  of the second auxiliary variable and 0 otherwise and so on. The population total  $T_x$  of this vector thus contains the population counts in each of the (marginal) categories of each of the three auxiliary variables. The construction of the weights for classical raking ratio estimation has traditionally involved the use of iterative proportional fitting (Brackstone and Rao 1979). Ireland and Kullback (1968) demonstrate that this method converges to a solution of the above optimisation problem.

The function  $G_{ML}(u)$  leads to an alternative 'maximum likelihood' version of raking adjustment, when  $x_i$  takes the same form, denoting indicator variables of categorical auxiliary variables. In this case, the objective function in (2.1) may be interpreted as a quantity which is proportional to minus a log likelihood in the case of simple random sampling with replacement (Brackstone and Rao 1979; Fuller 2002).

### 3. Asymptotic framework and nonresponse bias

We now consider the asymptotic properties of  $\hat{T}_y$  with respect to both the sampling design and the nonresponse mechanism. We assume that the latter is such that each unit in the population responds, if sampled, with probability  $q_i$ , where this probability is not dependent on the choice of the sample and different units respond independently. We consider an asymptotic framework defined in terms of sequences of finite populations and associated probability sampling designs and response mechanisms (Fuller 2009, section 1.3), with orders of magnitude terms expressed in terms of  $n = \sum_U \pi_i q_i$ , the expected number of responding units, and  $N$ , the population size. We assume there exist positive constants  $K_1, K_2$  and  $K_3$  such that  $K_1 < nN^{-1}d_i < K_2$  and  $K_3 < q_i$  for all  $i$ .

We shall suppose that Horvitz-Thompson estimators of means are consistent for the corresponding finite population



means and that central limit theorems hold (as expressed formally in the conditions of Theorem 1.3.9 of Fuller 2009). In particular, we assume that the sequences and the function  $F(\cdot)$  are such that there is a unique solution  $\lambda$  of

$$\sum_U q_i F(x'_i \lambda) x_i = T_x, \quad (3.1)$$

with

$$\hat{\lambda} = \lambda + O_p(n^{-0.5}), \quad (3.2)$$

and that

$$\hat{T}_y = \sum_U q_i F(x'_i \lambda) y_i + O_p(Nn^{-0.5}). \quad (3.3)$$

Déville and Särndal (1992) show that  $\lambda = 0$  under certain assumptions (their Result 2). However, their assumptions apply just to the distribution induced by the sampling design and include the requirement that  $N^{-1}(\hat{T}_{xd} - T_x) \rightarrow 0$  in probability. In the case of nonresponse, however, this requirement will often be implausible (*c.f.* Fuller 2002, page 15) and we do not require that  $\lambda$  be the zero vector.

A key assumption which we shall make is:

*Condition C:* there exists a vector  $\alpha$  such that  $F(x'_i \alpha) = q_i^{-1}$ .

If condition C holds then  $\alpha$  solves (3.1) and so  $\lambda = \alpha$ . It follows from (3.3) that  $\hat{T}_y$  is consistent for  $T_y$  for any choice of variable  $y$  if this condition holds. Thus, we may view condition C as a sufficient condition for the absence of (asymptotic) nonresponse bias. This property of Condition C has been discussed by Fuller, Loughlin and Baker (1994), Fuller (2009, page 284) and Särndal and Lundström (2005, Proposition 9.2) for the case when  $F$  is linear. Fuller (2002, page 15), Kott (2006) and Chang and Kott (2008) also consider estimating response probabilities using general models of the form  $q_i^{-1} = F(x'_i \alpha)$ .

To illustrate what might happen if condition C does not hold, suppose that  $x_i$  is just a scalar with  $x_i \equiv 1$ . Then the unique solution of (3.1) is  $\lambda = g(N/\sum_U q_i)$  and  $p\lim(\hat{T}_y) = N(\sum_U q_i y_i)/(\sum_U q_i)$ . Hence, the asymptotic nonresponse bias will only disappear for those survey variables which are 'uncorrelated' with the response probabilities  $q_i$ .

#### 4. Linearization variance estimation

We now proceed to consider the asymptotic variance of  $\hat{T}_y$  and its estimation. As in the previous section, the variance is defined with respect to the joint distribution induced by both sampling and nonresponse.

Note first that in general (and in particular for  $G_M(\cdot)$  and  $G_{ML}(\cdot)$ ), iteration is needed to solve the calibration equations. There does exist a literature (see Déville *et al.* 1993) which seeks to estimate the variance of  $\hat{T}_y$  after a finite

number of iterations. We follow instead the approach of Déville *et al.* (1993) and, for example, Binder and Thériège (1988) by approximating the variance of  $\hat{T}_y$  by the variance of the 'converged' estimator, *i.e.*, the hypothetical estimator arising from an infinite number of iterations, represented by  $\text{var}(\sum_s w_i y_i)$ , where the  $w_i$  are the 'converged' weights which solve the constrained optimisation problem in section 2.

A linearization variance estimator is obtained by approximating  $\text{var}(\sum_s w_i y_i)$  by  $\text{var}(\sum_s d_i z_i)$  for a 'linearized variable'  $z_i$  (Déville 1999). We now seek to construct this variable using a large sample argument. We first obtain an expression for  $\hat{\lambda}$ . A Taylor expansion of the left side of the calibration equations in (2.3) gives

$$\sum_i d_i F(x'_i \hat{\lambda}) x_i = \sum_s d_i F_i x_i + \sum_i d_i f(x'_i \lambda^*) x_i x'_i (\hat{\lambda} - \lambda),$$

where  $F_i = F(x'_i \lambda)$ ,  $\lambda^*$  is between  $\hat{\lambda}$  and  $\lambda$  and  $f(u) = dF(u)/du$  is assumed to exist. Assuming also continuity of  $f(\cdot)$ , the existence of  $\lim_{N \rightarrow \infty} N^{-1} \sum_U q_i f_i x_i x'_i$  and using (3.2), we have

$$N^{-1} \sum_i d_i F(x'_i \hat{\lambda}) x_i = N^{-1} \sum_s d_i F_i x_i + N^{-1} \sum_s d_i f_i x_i x'_i (\hat{\lambda} - \lambda) + o_p(n^{-0.5}), \quad (4.1)$$

where  $f_i = f(x'_i \lambda)$ . Then, assuming  $\lim_{N \rightarrow \infty} N^{-1} \sum_U q_i f_i x_i x'_i$  is non-singular and using (2.3), we obtain

$$\hat{\lambda} - \lambda = \left[ \sum_s d_i f_i x_i x'_i \right]^{-1} \left[ T_x - \sum_s d_i F_i x_i \right] + o_p(n^{-0.5}). \quad (4.2)$$

See Fuller (2009, proof of Theorem 1.3.9) for formal details of how (4.1) and (4.2) may be derived and the underlying regularity conditions. Note that to ensure  $\lim_{N \rightarrow \infty} N^{-1} \sum_U q_i f_i x_i x'_i$  is non-singular may require dropping redundant variables from  $x_i$  and possibly (as in Déville and Särndal 1992) modifying the estimator for samples with small probability that result in singularity of this matrix.

A similar argument involving the Taylor expansion of  $w_i$  in (2.2) about  $\lambda$  gives:

$$w_i = d_i [F_i + f_i x'_i (\hat{\lambda} - \lambda)] + o_p(Nn^{-1.5}). \quad (4.3)$$

Then, assuming the existence of necessary population moments so that the remainder term in (4.3) holds uniformly across  $i$  (Fuller 2009, Corollary 2.7.1.1.), we have

$$\begin{aligned} \hat{T}_y &\equiv \sum_s w_i y_i \\ &= \sum_i d_i \left[ F_i + f_i x'_i (\hat{\lambda} - \lambda) \right] y_i + o_p(Nn^{-0.5}) \end{aligned} \quad (4.4)$$



and hence from (4.2) and (4.4):

$$\hat{T}_y = \sum_s d_i F_i y_i + B \left[ T_x - \sum_s d_i F_i x_i \right] + o_p(Nn^{-0.5}), \quad (4.5)$$

where

$$B = \left[ \sum_s d_i f_i y_i x_i' \right] \left[ \sum_s d_i f_i x_i x_i' \right]^{-1}. \quad (4.6)$$

Note that  $F_i = f_i = 1$  under the assumptions of Deville and Särndal (1992) (since in this case  $\lambda = 0$  and it follows from the assumptions about  $G(\cdot)$  that  $F(0) = f(0) = 1$ ). Hence, under these assumptions, expression (4.5) corresponds to Result 5 of Deville and Särndal (1992), i.e., the generalized raking estimator is asymptotically equivalent to the GREG estimator. Therefore, the asymptotic variance of  $\hat{T}_y$  is the same as that of  $\sum_s d_i z_i$ , where  $z_i$  is the linearized variable:

$$z_i = F_i(y_i - \beta x_i), \quad (4.7)$$

and it is assumed that  $B$  converges to a finite limit matrix  $\beta$ . An alternative derivation of this expression is given by Demnati and Rao (2004, section 3.4).

For the purpose of linearization variance estimation,  $\hat{T}_y$  is treated as the linear estimator  $\sum_s d_i \hat{z}_i$ , where

$$\hat{z}_i = \hat{F}_i(y_i - \hat{B} x_i) \quad (4.8)$$

is treated as a fixed variable.

A number of choices of  $\hat{F}_i$  and  $\hat{B}$  have been discussed in the literature. Starting with  $\hat{F}_i$ , the natural choice implied by the above argument is  $\hat{F}_i = F(x_i' \hat{\lambda})$ . A simpler choice, however, would be to take  $\hat{F}_i = 1$ . Deville and Särndal (1992) note that, in their classical theory with  $\lambda = 0$ , these choices are asymptotically equivalent but they express a preference for the choice  $\hat{F}_i = F(x_i' \hat{\lambda})$ . In our setting with nonresponse and with  $\lambda = 0$  not necessarily holding, the second choice seems preferable and this is emphasized by Fuller (2002, page 15). Note that these two choices imply that  $\sum_s d_i \hat{z}_i$  either takes the form  $\sum w_i(y_i - \hat{B} x_i)$  when  $\hat{F}_i = F(x_i' \hat{\lambda})$  or  $\sum d_i(y_i - \hat{B} x_i)$  when  $\hat{F}_i = 1$ . We shall therefore refer to these choices as either  $w_i$ -weighted residuals or  $d_i$ -weighted residuals.

Regarding  $\hat{B}$ , it follows from our argument on the choices of  $\hat{F}_i$  that  $f_i$  in (4.2) should be replaced by  $\hat{f}_i = f(x_i' \hat{\lambda})$ , giving:

- (i)  $\hat{B} = [\sum_s d_i \hat{f}_i y_i x_i'] [\sum_s d_i \hat{f}_i x_i x_i']^{-1}$ , as also proposed by Demnati and Rao (2004).

Other choices are

- (ii)  $\hat{B} = \hat{B}_s$ , as in (2.4), as proposed by Deville *et al.* (1993).

- (iii)  $\hat{B} = [\sum_s w_i y_i x_i'] [\sum_s w_i x_i x_i']^{-1}$ , as proposed by Deville and Särndal (1992, equation 3.4), which

might be more practical to compute than  $\hat{B}_s$  for users of survey data files which include the  $w_i$  weights but not the  $d_i$  weights.

The extent to which these choices differ depends on the choice of  $G(\cdot)$  function. For the linear case  $f(u) = 1$  so that the estimators in (i) and (ii) are identical. In the case of classical raking adjustment,  $f(u) = F(u) = \exp(u)$  so that  $\hat{f}_i = \hat{F}_i$  and  $d_i \hat{f}_i = w_i$  and the estimators (i) and (iii) are identical. For the 'maximum likelihood' raking estimator we have  $F(u) = (1-u)^{-1}$  and  $f(u) = (1-u)^{-2}$  so that  $d_i \hat{f}_i = w_i^2/d_i$  and the three variance estimators are all distinct.

Having determined the form of  $\hat{z}_i$  in (4.8), the linearization variance estimator for  $\hat{T}_y$  is obtained by estimating the variance of the linear estimator  $\sum_s d_i \hat{z}_i$ , treating  $d_i$  and  $\hat{z}_i$  as fixed. In the case of a stratified multistage sampling design, assuming "with replacement" sampling of primary sampling units (PSUs) within strata, a standard estimator of the variance (e.g., Stukel *et al.* 1996) is:

$$\hat{V}(\hat{T}_y) = \sum_{h=1}^H \frac{n_h}{n_h - 1} \sum_{j=1}^{n_h} (z_{hj} - \bar{z}_h)^2 \quad (4.9)$$

where  $z_{hj} = \sum_k d_{hjk} \hat{z}_{hjk}$ ,  $\bar{z}_h = \sum_j z_{hj}/n_h$  and  $\hat{z}_{hjk}$  is the value of the variable defined in (4.8) for the  $k^{\text{th}}$  individual within the  $j^{\text{th}}$  selected PSU in stratum  $h$ . This estimator remains appropriate in the presence of nonresponse if individual response in each PSU is independent of response in all other PSUs and if at least one individual is observed in each selected PSU (Fuller *et al.* 1994, page 78).

## 5. Simulation studies

In order to compare the performance of the weighted estimators and their corresponding variance estimators, two simulation studies were undertaken by constructing artificial populations using data from the British Labour Force Survey (LFS) and the German Sample Survey of Income and Expenditure (SIE). In each case,  $R = 1,000$  samples were generated from these populations by first sampling, in a way designed to mimic the real sampling scheme after some simplification, and then removing nonresponding cases according to two nonresponse models. The first assumes multiplicative nonresponse which, from Condition C in section 3, might be expected to lead to least bias for the raking ratio method. The second model assumed additive nonresponse, which might be expected to lead to least bias for the GREG estimator.

For each of the  $R$  samples, point estimates of parameters were calculated using the different generalized raking methods presented in section 2 and variance estimates were calculated using the different linearization methods presented in section 4. The properties of the estimators were then summarised.

## 5.1 Study based on the British Labour Force Survey

The first study was based upon data from the March-May 1998 quarter of the British LFS, a survey of persons living in private households in Britain, designed to provide information on the British labour market and carried out by the Office for National Statistics (ONS). The sample of approximately 58,000 households was treated as an artificial population. Repeated samples were drawn from this population in a way intended to mimic the design used for the LFS (ONS 1998, Section 3). Each sample consisted of 1,211 households selected by stratified simple random sampling with proportional allocation across 19 strata, defined by region of residence. These regions were designed to mimic interviewer areas which defined strata in the LFS. In the LFS all individuals in a sampled household are interviewed if possible. In this simulation study, all the respondents in a sample household were retained, except those aged under 16, who are not relevant for the estimates of interest.

The following two nonresponse models, based upon results of a study of Foster (1998), were used to determine whether sampled individuals responded.

*Multiplicative Nonresponse Model:*

$$q_i^{-1} = 1.15 \times \begin{array}{l} 1.17 \text{ (if London)} \\ 1.13 \text{ (if aged under 35)} \\ 1.1 \text{ (if female)} \end{array}$$

*Additive Nonresponse Model:*

$$q_i^{-1} = 1.15 + \begin{array}{l} 0.20 \text{ (if London)} \\ 0.15 \text{ (if aged under 35)} \\ 0.10 \text{ (if female)} \end{array}$$

where  $q_i$  is the response probability defined at the beginning of section 3 and the form of the model is chosen to satisfy Condition C.

Three parameters of interest are defined for the artificial population: the total number of persons unemployed, employed or inactive in the workforce. Weights were constructed for responding individuals, with calibration totals consisting of population counts in the categories of three categorical auxiliary variables and with Horvitz-Thompson initial weights  $d_i$ , as in section 2. The choice of auxiliary variables was designed to mimic those used in the LFS. However, because of the reduced scale of our artificial population and the consequent smaller numbers of individuals within strata, we simplified the LFS calibration variables to the following three categorical factors, defining 83 control totals:

- a cross-classification of sex by 10 age groups (consisting of single years for those between 16 and 24 and a separate age group for 25 or older) with 20 categories;
- a cross-classification of region (Northern England; London and South East; Midlands and East Anglia; Scotland) by sex by age in 15-year age groups (16-29, 30-44, 45-59, 60-75 and 75 or older) with 40 categories.

## 5.2 Study based on the German sample Survey of Income and Expenditure

Our second study is based on the 1998 German Survey of Income and Expenditure (SIE), a national household survey conducted every 5 years by the Federal Statistical Office, to provide information about the economic and social situation of households, especially regarding the distribution of income and expenditure (Muennich and Schulrle 2003). We used data from a synthetic population of 64,326 households, created to represent 20% of all households from the Bremen region, excluding those with a monthly household net income of DM 35,000 or above (DM denotes the currency of German marks). A quota sampling design was employed for this survey and we have not attempted to mimic this design. Instead, our simulation study employs simple random sampling together with nonresponse. Repeated simple random samples of 1,340 households were drawn from the artificial population, representing a sampling fraction of about 1/48. Nonresponse models were constructed using the results of studies of similar surveys in Great Britain: the Family Expenditure Survey and the National Food Survey (Foster 1998). For each selected sample, the subset of responding households was determined by the following nonresponse models:

*Multiplicative Model:*

$$q_i^{-1} = 1.44 \times \begin{array}{l} 1.09 \text{ (if self-employed)} \\ 1.03 \text{ (if unemployed)} \\ 0.97 \text{ (if employed)} \\ 1.16 \text{ (if no children in the household)} \end{array}$$

*Additive Model:*

$$q_i^{-1} = 1.44 + \begin{array}{l} 0.13 \text{ (if self-employed)} \\ 0.04 \text{ (if unemployed)} \\ - 0.04 \text{ (if employed)} \\ + 0.23 \text{ (if no children in the household)} \end{array}$$

The parameters of interest are the total household net income per quarter and the total household expenditure per quarter, computed from the finite artificial population.

As for the LFS study, each sampled household was assigned a weight. In the actual SIE the weights are constructed using essentially the maximum likelihood raking method by adjusting the sample data simultaneously to the marginal

distributions of several characteristics, such as household type, social economic status of the reference person, household net income class and region (land). We try to mimic this adjustment, as far as possible, in our study. However, as for the LFS, because of the problem of strata with small numbers of households we simplify the SIE calibration variables to the following three categorical factors:

- household type with 7 categories
  - mother/father alone + 1 child,
  - mother/father alone + 2 or more children,
  - couple with 1 child – spouse employed,
  - couple with 1 child – spouse unemployed,
  - couple with 2 or more children – spouse employed,
  - couple with 2 or more children – spouse unemployed,
  - other.
- social status of the reference person with 5 categories
  - self-employed,
  - civil servant or military,
  - employee,
  - worker,
  - unemployed, pensioner, student or other.
- household net income per quarter with 3 categories
  - 0-5,000 DM,
  - 5-7,000 DM,
  - 7-35,000 DM.

## 6. Results

### 6.1 Properties of point estimators

Table 6.1 presents the properties of the point estimators of total unemployed in the LFS study for different

calibration methods and alternative assumptions about nonresponse. The properties are assessed following usual practice in simulation studies. For example, the bias in Table 6.1 is obtained from  $\hat{B}(\hat{T}_y) = \hat{E}(\hat{T}_y) - T_y$ , where  $\hat{E}(\hat{T}_y) = 1/R \sum_{r=1}^R \hat{T}_{y_r}$ ,  $\hat{T}_{y_r}$  is the value of  $\hat{T}_y$  for sample  $r$  and  $R$  is the number of simulated samples. We observe from this table that the standard error remains virtually constant across alternative raking methods for a given nonresponse model. Nonresponse leads to an increase in the standard error across all estimators as expected (since the sample size is reduced). The table does show evidence of nonresponse bias, which is of a similar order for each of the raking methods. We do not find that this bias is least when the estimator matches the nonresponse model (*i.e.*, the GREG estimator for additive response and the raking estimator for multiplicative response) as we might have expected. Perhaps this is because the covariates used in the nonresponse models (*e.g.*, the aged 35+ variable) are not all included in the calibrating variables. Nevertheless, the nonresponse bias is small in the sense that the root mean square error is very similar to the standard error in each case. Under nonresponse, the GREG calibration method generates some negative weights whereas this is avoided by the two raking methods, as expected. A greater number of very large weights are observed, however, for the 'maximum likelihood' raking estimator.

Corresponding results for the SIE data are presented in Table 6.2. The pattern of results is broadly similar, although there is now no evidence of significant nonresponse bias (*i.e.*, the observed bias could be explained by simulation variation). The standard errors and root mean square errors also remain virtually constant across weighting methods for a given nonresponse model.

**Table 6.1**  
Simulation properties of point estimators of total unemployed using data from LFS with R = 1,000

Nonresponse Model/Point Estimator	Bias (simulation standard error)	Standard Error	Root Mean Square Error	Number of Negative Weights <sup>1</sup>	Number of Very Large Weights <sup>1,2</sup>
<i>Complete Response:</i>					
GREG	7.6 (14.3)	452.8	452.8	0	0
Classical Raking	8.3 (14.3)	452.8	452.9	0	0
'ML' Raking	9.0 (14.3)	453.3	453.4	0	1
<i>Multiplicative nonresponse:</i>					
GREG	-45.6 (15.8)	498.3	500.3	4	1
Classical Raking	-42.1 (15.8)	498.8	500.6	0	2
'ML' Raking	-39.7 (15.8)	499.4	501.0	0	7
<i>Additive nonresponse:</i>					
GREG	-37.3 (15.7)	497.4	498.8	5	1
Classical Raking	-34.7 (15.7)	497.5	498.7	0	3
'ML' Raking	-32.4 (15.8)	498.1	499.1	0	7

<sup>1</sup> the number of such weights across all sample units and all 1000 samples.

<sup>2</sup> the number of weights more than 10 times the corresponding design weight.



Table 6.2

Simulation properties of point estimators of total income using data from SIE with R = 1,000

Nonresponse Model/Point Estimator	Bias (simulation standard error)	Standard Error	Root Mean Square Error	Number of Negatives Weights	Number of Very Large Weights
<i>Complete Response:</i>					
GREG	-172.2 (331.3)	10,477.3	10,478.7	0	0
Classical Raking	-170.6 (331.5)	10,484.1	10,485.8	0	0
'ML' Raking	-169.8 (331.8)	10,491.5	10,492.9	0	0
<i>Multiplicative nonresponse:</i>					
GREG	-495.7 (429.7)	13,586.8	13,595.8	0	0
Classical Raking	-493.8 (429.6)	13,584.6	13,593.5	0	0
'ML' Raking	-463.5 (429.5)	13,582.8	13,590.7	0	0
<i>Additive nonresponse:</i>					
GREG	-473.2 (430.5)	13,614.8	13,623.0	0	0
Classical Raking	-469.4 (430.5)	13,612.9	13,621.0	0	0
'ML' Raking	-439.5 (430.5)	13,613.5	13,620.6	0	0

## 6.2 Properties of variance estimators

The properties of the different estimators of the variances of the point estimators of the total unemployed from the LFS are shown in the Table 6.3 (the 'standard error estimate' in the table refers to the square root of the variance estimate). We make a number of observations:

- weighting the residuals by  $w_i$  rather than by  $d_i$  reduces the bias and root mean squared error of the standard error estimator. The bias arising from the use of  $d_i$  weighted residuals in the case of nonresponse is particularly important (as noted by Fuller 2002) but there are also non-negligible reductions of bias even in the complete response case.
- The choice of weight used in  $\hat{B}$  for the calculation of residuals seems to have little impact.
- For a given nonresponse setting and choice of weighting the residuals, there is little difference in the results for the different choices of point estimator.

The results in Table 6.3 are extended in Table 6.4 to consider relative bias of the standard error estimators, rather than their absolute bias, and to consider two additional parameters: total numbers employed and inactive. We see again that the relative bias arising from using  $d_i$  weighted

residuals can be substantial in the presence of nonresponse, over 20% in several cases, and that this is reduced using the  $w_i$  weighted residuals. Again, little change is observed in the percent relative bias of the standard error estimators when different choices of weights are used in the calculation of  $\hat{B}$  for the residuals.

Corresponding results for the SIE data when estimating total income are shown in Table 6.5. Again, the pattern of results is broadly similar to that for the LFS data in Table 6.3. For the complete response case, the use of  $w_i$  weighted residuals rather than  $d_i$  weighted residuals leads to modest improvement in bias and RMSE of the standard error estimators. For the nonresponse cases the improvements are considerable. Little change in the standard error estimators is observed when modifying the choice of weight used to compute the estimated regression coefficients. The results in Table 6.5 are extended in Table 6.6 to consider relative bias of the standard error estimators, rather than their absolute bias, and to consider one additional parameter: total expenditure per quarter. We see again that the relative bias arising from using  $d_i$  weighted residuals can be substantial in the presence of nonresponse, over 35% in all cases, and that this is reduced using the  $w_i$  weighted residuals, for which the relative bias never exceeds about 3%.

**Table 6.3**  
**Properties of variance estimators when estimating total unemployed from the LFS (R = 1,000)**

Weighting Method	<i>w</i> - or <i>d</i> - weighted residuals <sup>1</sup>	weight used for $\hat{B}$ in residual <sup>1</sup>	Mean of Standard Error Estimator	Bias of SE Estimator (simulation s.e.)	RMSE of SE Estimator	Coverage <sup>2</sup> of Confidence Interval (%)
<i>Complete Response:</i>						
GREG	<i>d</i>	<i>d</i>	433.9	-18.8 (0.9)	33.4	93.5
	<i>d</i>	<i>w</i>	434.3	-18.5 (0.9)	33.3	93.5
	<i>w</i>	<i>d</i>	442.8	-10.0 (1.0)	31.9	93.8
	<i>w</i>	<i>w</i>	441.9	-10.8 (1.0)	32.0	93.7
Classical Raking	<i>d</i>	<i>d</i>	433.9	-18.8 (0.9)	33.4	93.5
	<i>d</i>	<i>w</i>	434.2	-18.5 (0.9)	33.3	93.5
	<i>w</i>	<i>d</i>	443.0	-9.8 (1.0)	32.0	93.8
	<i>w</i>	<i>w</i>	442.0	-10.7 (1.0)	32.0	93.8
'ML' Raking	<i>d</i>	<i>d</i>	433.9	-19.4 (0.9)	33.7	93.5
	<i>d</i>	<i>w</i>	434.3	-19.1 (0.9)	33.6	93.5
	<i>d</i>	<i>df</i>	435.4	-17.9 (0.9)	33.0	93.5
	<i>w</i>	<i>d</i>	443.7	-9.6 (1.0)	32.5	93.7
	<i>w</i>	<i>w</i>	442.3	-11.1 (1.0)	32.4	93.7
	<i>w</i>	<i>df</i>	441.6	-11.8 (1.0)	32.3	93.7
<i>Multiplicative nonresponse:</i>						
GREG	<i>d</i>	<i>d</i>	385.7	-112.6 (0.9)	116.0	85.8
	<i>d</i>	<i>w</i>	386.1	-112.1 (0.9)	115.5	85.8
	<i>w</i>	<i>d</i>	489.5	-8.8 (1.2)	39.2	94.2
	<i>w</i>	<i>w</i>	487.8	-10.4 (1.2)	39.2	94.2
Classical Raking	<i>d</i>	<i>d</i>	385.7	-113.1 (0.9)	116.5	85.7
	<i>d</i>	<i>w</i>	386.1	-112.7 (0.9)	116.1	85.7
	<i>w</i>	<i>d</i>	490.3	-8.5 (1.2)	39.6	94.3
	<i>w</i>	<i>w</i>	488.4	-10.4 (1.2)	39.5	94.1
'ML' Raking	<i>d</i>	<i>d</i>	385.7	-113.7 (0.9)	117.1	85.4
	<i>d</i>	<i>w</i>	386.2	-113.2 (0.9)	116.6	85.6
	<i>d</i>	<i>df</i>	387.8	-111.6 (0.9)	115.0	85.8
	<i>w</i>	<i>d</i>	491.9	-7.5 (1.3)	40.4	94.2
	<i>w</i>	<i>w</i>	488.9	-10.5 (1.2)	39.9	94.0
	<i>w</i>	<i>df</i>	487.5	-11.9 (1.2)	39.8	94.0
<i>Additive nonresponse:</i>						
GREG	<i>d</i>	<i>d</i>	386.5	-110.9 (0.9)	114.4	86.0
	<i>d</i>	<i>w</i>	387.0	-110.5 (0.9)	113.9	86.0
	<i>w</i>	<i>d</i>	489.3	-8.2 (1.2)	39.0	94.6
	<i>w</i>	<i>w</i>	487.6	-9.8 (1.2)	39.0	94.6
Classical Raking	<i>d</i>	<i>d</i>	386.5	-111.0 (0.9)	114.4	85.8
	<i>d</i>	<i>w</i>	387.0	-110.6 (0.9)	114.0	85.8
	<i>w</i>	<i>d</i>	490.1	-7.4 (1.2)	39.2	94.7
	<i>w</i>	<i>w</i>	488.1	-9.4 (1.2)	39.1	94.6
'ML' Raking	<i>d</i>	<i>d</i>	386.5	-111.6 (0.9)	115.0	85.6
	<i>d</i>	<i>w</i>	387.0	-111.1 (0.9)	114.6	85.6
	<i>d</i>	<i>df</i>	388.6	-109.5 (0.9)	113.0	85.9
	<i>w</i>	<i>d</i>	491.6	-6.5 (1.3)	40.0	94.7
	<i>w</i>	<i>w</i>	488.6	-9.5 (1.2)	39.5	94.6
	<i>w</i>	<i>df</i>	487.3	-10.8 (1.2)	39.4	94.6

<sup>1</sup> see text following equation (4.8), where choices *df*, *d* and *w* correspond to  $\hat{B}$  in (i), (ii) and (iii) respectively.

<sup>2</sup> percentage of 95% normal-theory confidence intervals containing true value.

Table 6.4

Relative bias (%) of standard error estimators of unemployed, employed and inactive totals from LFS (R = 1,000)

Weighting Method	<i>w</i> - or <i>d</i> -weighted residuals <sup>1</sup>	weight used for $\hat{B}$ in residual <sup>1</sup>	Relative Bias of Standard Error Estimator		
			Unemployed	Employed	Inactive
<i>Complete Response:</i>					
GREG	<i>d</i>	<i>d</i>	-4.2	-3.4	0.5
	<i>d</i>	<i>w</i>	-4.1	-3.3	0.6
	<i>w</i>	<i>d</i>	-2.2	-2.2	1.9
	<i>w</i>	<i>w</i>	-2.4	-2.3	1.7
Classical Raking	<i>d</i>	<i>d</i>	-4.2	-3.3	0.7
	<i>d</i>	<i>w</i>	-4.1	-3.2	0.8
	<i>w</i>	<i>d</i>	-2.2	-2.1	2.1
	<i>w</i>	<i>w</i>	-2.4	-2.2	1.9
'ML' Raking	<i>d</i>	<i>d</i>	-4.3	-3.3	0.7
	<i>d</i>	<i>w</i>	-4.2	-3.3	0.8
	<i>d</i>	<i>df</i>	-4.0	-3.1	1.1
	<i>w</i>	<i>d</i>	-2.1	-2.0	2.3
	<i>w</i>	<i>w</i>	-2.4	-2.2	1.9
	<i>w</i>	<i>df</i>	-2.6	-2.3	1.8
<i>Multiplicative nonresponse:</i>					
GREG	<i>d</i>	<i>d</i>	-22.6	-22.3	-18.2
	<i>d</i>	<i>w</i>	-22.5	-22.2	-18.1
	<i>w</i>	<i>d</i>	-1.8	-3.3	1.8
	<i>w</i>	<i>w</i>	-2.1	-3.5	1.5
Classical Raking	<i>d</i>	<i>d</i>	-22.7	-30.6	-18.4
	<i>d</i>	<i>w</i>	-22.6	-30.5	-18.3
	<i>w</i>	<i>d</i>	-1.7	-13.5	1.7
	<i>w</i>	<i>w</i>	-2.1	-13.7	1.3
'ML' Raking	<i>d</i>	<i>d</i>	-22.8	-22.0	-18.4
	<i>d</i>	<i>w</i>	-22.7	-21.9	-18.3
	<i>d</i>	<i>df</i>	-22.3	-21.7	-17.9
	<i>w</i>	<i>d</i>	-1.5	-2.7	1.9
	<i>w</i>	<i>w</i>	-2.1	-3.1	1.3
	<i>w</i>	<i>df</i>	-2.4	-3.3	1.1
<i>Additive nonresponse:</i>					
GREG	<i>d</i>	<i>d</i>	-22.3	-21.8	-18.5
	<i>d</i>	<i>w</i>	-22.2	-21.7	-18.4
	<i>w</i>	<i>d</i>	-1.6	-2.9	1.1
	<i>w</i>	<i>w</i>	-2.0	-3.1	0.8
Classical Raking	<i>d</i>	<i>d</i>	-22.3	-30.2	-18.0
	<i>d</i>	<i>w</i>	-22.2	-30.1	-17.9
	<i>w</i>	<i>d</i>	-1.5	-13.3	1.8
	<i>w</i>	<i>w</i>	-1.9	-13.5	1.4
'ML' Raking	<i>d</i>	<i>d</i>	-22.4	-21.6	-18.0
	<i>d</i>	<i>w</i>	-22.3	-21.5	-17.9
	<i>d</i>	<i>df</i>	-22.0	-21.3	-17.6
	<i>w</i>	<i>d</i>	-1.3	-2.4	2.0
	<i>w</i>	<i>w</i>	-1.9	-2.8	1.5
	<i>w</i>	<i>df</i>	-2.2	-3.0	1.3

<sup>1</sup> see text following equation (4.8), where *df*, *d* and *w* correspond to  $\hat{B}$  in (i), (ii) and (iii) respectively.



**Table 6.5**  
**Properties of variance estimators when estimating total income from the SIE (R = 1,000)**

Weighting Method	w- or d-weighted residuals <sup>1</sup>	weight used for $\hat{B}$ in residual <sup>1</sup>	Mean of Standard Error Estimator	Bias of SE Estimator (s.e.)	RMSE of SE Estimator	Coverage <sup>2</sup> of Confidence Interval (%)
<i>Complete Response:</i>						
GREG	d	d	10,338.8	-138.5 (6.9)	259.0	93.8
	d	w	10,339.2	-138.2 (6.9)	258.8	93.8
	w	d	10,377.9	-99.5 (6.9)	240.0	94.1
	w	w	10,376.8	-100.5 (6.9)	240.3	94.1
Classical Raking	d	d	10,338.8	-145.3 (6.9)	262.7	93.8
	d	w	10,339.2	-144.9 (6.9)	262.5	93.8
	w	d	10,370.0	-106.1 (6.9)	243.1	94.0
	w	w	10,376.9	-107.2 (6.9)	243.5	94.0
'ML' Raking	d	d	10,338.8	-152.7 (6.9)	266.9	93.9
	d	w	10,339.2	-152.4 (6.9)	266.7	93.9
	d	df	10,340.3	-151.3 (6.9)	266.1	94.0
	w	d	10,378.3	-113.2 (6.9)	246.5	94.0
	w	w	10,377.1	-114.4 (6.9)	247.0	94.0
	w	df	10,376.7	-114.8 (6.9)	247.2	94.0
<i>Multiplicative nonresponse:</i>						
GREG	d	d	8,104.7	-5,482.1 (7.4)	5,487.1	75.8
	d	w	8,105.5	-5,481.3 (7.4)	5,486.3	75.8
	w	d	13,214.5	-372.3 (12.8)	549.7	94.5
	w	w	13,210.9	-375.9 (12.8)	551.7	94.5
Classical Raking	d	d	8,104.7	-5,479.8 (7.4)	5,484.9	75.8
	d	w	8,105.5	-5,479.1 (7.4)	5,484.1	75.8
	w	d	13,214.1	-370.4 (12.8)	549.4	94.5
	w	w	13,210.4	-374.2 (12.8)	551.5	94.5
'ML' Raking	d	d	8,104.7	-5,478.1 (7.4)	5,483.1	75.8
	d	w	8,105.5	-5,477.3 (7.4)	5,482.3	75.8
	d	df	8,108.1	-5,474.7 (7.4)	5,479.7	75.9
	w	d	13,215.2	-367.6 (12.9)	549.4	94.5
	w	w	13,210.6	-372.2 (12.9)	551.6	94.5
	w	df	13,208.9	-373.9 (12.9)	552.3	94.5
<i>Additive nonresponse:</i>						
GREG	d	d	8,106.3	-5,508.5 (7.4)	5,513.5	75.6
	d	w	8,107.1	-5,507.7 (7.4)	5,512.7	75.6
	w	d	13,207.9	-407.0 (12.8)	573.8	94.3
	w	w	13,204.3	-410.5 (12.8)	575.9	94.3
Classical Raking	d	d	8,106.3	-5,506.6 (7.4)	5,511.6	75.7
	d	w	8,107.1	-5,505.9 (7.4)	5,510.9	75.7
	w	d	13,207.7	-405.3 (12.8)	573.6	94.1
	w	w	13,203.9	-409.0 (12.8)	575.8	94.1
'ML' Raking	d	d	8,106.3	-5,507.2 (7.4)	5,512.2	75.9
	d	w	8,107.1	-5,506.4 (7.4)	5,511.4	75.9
	d	df	8,109.7	-5,503.8 (7.4)	5,508.8	75.9
	w	d	13,208.9	-404.6 (12.9)	574.8	94.1
	w	w	13,204.2	-409.2 (12.9)	577.3	94.1
	w	df	13,202.5	-411.0 (12.9)	578.1	94.1

<sup>1</sup>see text following equation (4.8), where choices *df*, *d* and *w* correspond to  $\hat{B}$  in (i), (ii) and (iii) respectively.

<sup>2</sup>percentage of 95% normal-theory confidence intervals containing true value.

**Table 6.6**  
**Relative bias (%) of variance estimators of expenditure and income totals from SIE (R = 1,000)**

Weighting Method	<i>w</i> - or <i>d</i> -weighted residuals <sup>1</sup>	weight used for $\hat{B}$ in residual <sup>1</sup>	Relative Bias of Standard Error Estimator	
			Expenditure	Income
<i>Complete Response:</i>				
GREG	<i>d</i>	<i>d</i>	0.7	-1.3
	<i>d</i>	<i>w</i>	0.7	-1.3
	<i>w</i>	<i>d</i>	1.3	-1.0
	<i>w</i>	<i>w</i>	1.3	-1.0
Classical Raking	<i>d</i>	<i>d</i>	0.7	-1.4
	<i>d</i>	<i>w</i>	0.7	-1.4
	<i>w</i>	<i>d</i>	1.2	-1.0
	<i>w</i>	<i>w</i>	1.2	-1.0
'ML' Raking	<i>d</i>	<i>d</i>	0.6	-1.5
	<i>d</i>	<i>w</i>	0.6	-1.5
	<i>d</i>	<i>df</i>	0.6	-1.4
	<i>w</i>	<i>d</i>	1.2	-1.1
	<i>w</i>	<i>w</i>	1.2	-1.1
	<i>w</i>	<i>df</i>	1.2	-1.1
<i>Multiplicative nonresponse:</i>				
GREG	<i>d</i>	<i>d</i>	-38.2	-40.4
	<i>d</i>	<i>w</i>	-38.2	-40.3
	<i>w</i>	<i>d</i>	-0.3	-2.7
	<i>w</i>	<i>w</i>	-0.3	-2.8
Classical Raking	<i>d</i>	<i>d</i>	-38.2	-40.3
	<i>d</i>	<i>w</i>	-38.2	-40.3
	<i>w</i>	<i>d</i>	-0.3	-2.7
	<i>w</i>	<i>w</i>	-0.3	-2.8
'ML' Raking	<i>d</i>	<i>d</i>	-38.2	-40.3
	<i>d</i>	<i>w</i>	-38.2	-40.3
	<i>d</i>	<i>df</i>	-38.2	-40.3
	<i>w</i>	<i>d</i>	-0.3	-2.7
	<i>w</i>	<i>w</i>	-0.3	-2.7
	<i>w</i>	<i>df</i>	-0.4	-2.8
<i>Additive nonresponse:</i>				
GREG	<i>d</i>	<i>d</i>	-38.1	-40.5
	<i>d</i>	<i>w</i>	-38.1	-40.5
	<i>w</i>	<i>d</i>	-0.2	-3.0
	<i>w</i>	<i>w</i>	-0.2	-3.0
Classical Raking	<i>d</i>	<i>d</i>	-38.1	-40.5
	<i>d</i>	<i>w</i>	-38.1	-40.5
	<i>w</i>	<i>d</i>	-0.2	-3.0
	<i>w</i>	<i>w</i>	-0.2	-3.0
'ML' Raking	<i>d</i>	<i>d</i>	-38.2	-40.5
	<i>d</i>	<i>w</i>	-38.2	-40.5
	<i>d</i>	<i>df</i>	-38.1	-40.4
	<i>w</i>	<i>d</i>	-0.2	-3.0
	<i>w</i>	<i>w</i>	-0.3	-3.0
	<i>w</i>	<i>df</i>	-0.3	-3.0

<sup>1</sup> see text following equation (4.8), where *df*, *d* and *w* correspond to  $\hat{B}$  in (i), (ii) and (iii) respectively.

## 7. Conclusions

The simulation study showed little difference between the bias or variance properties of the three calibration estimators considered: the GREG estimator, the classical raking estimator and the maximum likelihood raking estimator. Some small differences in the distribution of extreme weights were observed: the maximum likelihood raking estimator had the most very large weights and the GREG estimator was the only one with a few negative weights.

Amongst the variance estimators, the main finding was the contrast between the approach which weights residuals by the design weight and that which weights them by the calibrated weight. It was found that the latter variance estimator always had smaller bias and that this effect was very marked in the presence of nonresponse, when the former estimator could be severely biased. The bias of the latter estimator was generally small and the coverage level of the associated confidence intervals was generally close to the nominal coverage.

Alternative ways of weighting the observations in constructing the regression coefficients, when calculating the residuals in the linearization variance estimator, were considered but little effect was observed and there was no evidence that this choice is important in practice.

In general, the findings for the categorical variables in the British Labour Force Survey were remarkably similar to the findings for the continuous variables in the German Income and Expenditure survey.

## Acknowledgements

Comments from two referees helped improve this paper significantly. We are grateful to the Office for National Statistics for making the Labour Force Survey data available and to Ralf Münnich and colleagues on the DACSEIS project (<http://www.dacseis.de/>) for providing the synthetic population based on the German Survey of Income and Expenditure. This research was supported by the Economic and Social Research Council.

## References

- Binder, D.A., and Théberge, A. (1988). Estimating the variance of raking ratio estimators. *Canadian Journal of Statistics*, 16, Supp. 47-55.
- Brackstone, G.J., and Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Series C, 41, 97-114.
- Chang, T., and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.
- Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30, 17-34.
- Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-82.
- Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-20.
- Foster, K. (1998). Evaluating nonresponse on household surveys. *GSS Methodology Series*, 8, Office for National Statistics. London.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.
- Fuller, W.A. (2009). *Sampling Statistics*. Hoboken: Wiley.
- Fuller, W.A., Loughlin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.
- Ireland, C.T., and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.
- Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.
- Kalton, G., and Maligalig, D.S. (1991). A comparison of methods for weighting adjustment for nonresponse. *Proceedings of the US Bureau of the Census 1991 Annual Research Conference*, 409-428.
- Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.
- Muennich, R., and Schulrle, J. (2003). Monte Carlo simulation study of European surveys, Workpackage 3, Deliverables 3.1 and 3.2. DACSEIS project. Available at <http://www.uni-trier.de/index.php?id=29730>.
- Office for National Statistics (1998). *Labour Force Survey User Guide, Volume 1: Background and Methodology*. London.
- Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester, England.
- Stukel, D.M., Hidioglou, M.A. and Särndal, C.-E. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.



# Linearization variance estimators for model parameters from complex survey data

Abdellatif Demnati and J.N.K. Rao<sup>1</sup>

## Abstract

Taylor linearization methods are often used to obtain variance estimators for calibration estimators of totals and nonlinear finite population (or census) parameters, such as ratios, regression and correlation coefficients, which can be expressed as smooth functions of totals. Taylor linearization is generally applicable to any sampling design, but it can lead to multiple variance estimators that are asymptotically design unbiased under repeated sampling. The choice among the variance estimators requires other considerations such as (i) approximate unbiasedness for the model variance of the estimator under an assumed model, and (ii) validity under a conditional repeated sampling framework. Demnati and Rao (2004) proposed a unified approach to deriving Taylor linearization variance estimators that leads directly to a unique variance estimator that satisfies the above considerations for general designs. When analyzing survey data, finite populations are often assumed to be generated from super-population models, and analytical inferences on model parameters are of interest. If the sampling fractions are small, then the sampling variance captures almost the entire variation generated by the design and model random processes. However, when the sampling fractions are not negligible, the model variance should be taken into account in order to construct valid inferences on model parameters under the combined process of generating the finite population from the assumed super-population model and the selection of the sample according to the specified sampling design. In this paper, we obtain an estimator of the total variance, using the Demnati-Rao approach, when the characteristics of interest are assumed to be random variables generated from a super-population model. We illustrate the method using ratio estimators and estimators defined as solutions to calibration weighted estimating equations. Simulation results on the performance of the proposed variance estimator for model parameters are also presented.

Key Words: Calibration; Ratio estimators; Total variance; Logistic regression; Weighted estimating equations.

## 1. Introduction

In survey sampling, estimation of a finite population total  $Y = \sum_{k=1}^N y_k \equiv Y(y)$  is often of interest, where  $N$  is the size of the finite population. For a general sampling design with positive inclusion probabilities  $\pi_k$ , a customary design unbiased estimator of the total  $Y$  is given by  $\hat{Y} = \sum_{i \in s} y_i / \pi_i \equiv \sum_{k=1}^N d_k(s) y_k$ , where  $s$  is a sample,  $d_k(s) = a_k(s) / \pi_k$  are the design weights with  $a_k(s) = 1$  if  $k \in s$  and  $a_k(s) = 0$  otherwise. We use operator notation and write  $\hat{Y}(z) = \sum_{k=1}^N d_k(s) z_k$  so that  $\hat{Y} = \hat{Y}(y)$ . Henceforth, all the sums are considered on the whole population and hence write  $\sum_{k=1}^N y_k = \sum y_k$  and  $\hat{Y}(z) = \sum d_k(s) z_k$ , to simplify the notation. Again, using the operator notation, we denote an unbiased estimator of the variance of  $\hat{Y}(z)$  as a quadratic function,  $\mathfrak{V}(z)$ , in the  $z_k$ 's.

More complex estimators of a total  $Y$  based on known population auxiliary information, such as ratio and regression estimators, and estimators of more complex parameters obtained as solutions to sample weighted estimating equations, such as estimators of "census" logistic regression coefficients, are also often used in practice. Estimators that can be expressed as a general functional  $T(\hat{M})$  have also been studied, where  $\hat{M}$  denotes a measure that allocates the weight  $d_k(s)$  to  $y_k$ ;

for example,  $T(\hat{M}) = \int x d\hat{M}(x) = \sum d_k(s) y_k$  if the population parameter is the total  $T(M) = \int x dM(x) = Y$ , where the measure  $M$  allocates a unit mass to each  $y_k$  (Deville 1999). Large-sample estimation of the variance of such complex estimators,  $\hat{\theta}$  say, has received considerable attention in the literature. In particular, Taylor linearization methods of estimating the variance of  $\hat{\theta}$  are generally applicable to any sampling design that permits an unbiased variance estimator  $\mathfrak{V}(z)$  of  $\hat{Y}(z)$ . Binder (1983) studied estimators  $\hat{\theta}$  that are solutions to weighted estimating equations and applied Taylor linearization to obtain a variance estimator that can be expressed as  $\mathfrak{V}(\tilde{z})$ , where the linearized variable  $\tilde{z}_k$  depends on unknown parameters, and  $\tilde{z}_k$  is replaced by an estimator  $z_k$  that may be based on the substitution method. Deville (1999) derived a Taylor linearization variance estimator of the functional  $T(\hat{M})$  as  $\mathfrak{V}(\tilde{z})$ , where  $\tilde{z}_k = I_T(M; y_k)$  denotes the influence function of  $T$  at  $y_k$ , and then replaced  $\tilde{z}_k$  by the sample estimator  $z_{k1} = I_T(\hat{M}; y_k)$ . For example, when  $\hat{\theta}$  is the ratio estimator  $(\hat{Y} / \hat{X})X = \hat{R}X$  of the total  $Y$ , where  $\hat{X} = \hat{Y}(x)$  and  $X = Y(x)$  is the known total of an auxiliary variable  $x$ , we get  $\tilde{z}_k = y_k - Rx_k$  and  $z_{k1} = y_k - \hat{R}x_k$ . However,  $z_k = (X / \hat{X})(y_k - \hat{R}x_k)$  is also a candidate to estimate  $\tilde{z}_k$  and the resulting  $\mathfrak{V}(z)$  is often preferred over  $\mathfrak{V}(z_1)$ ; see Demnati and Rao (2004). Thus the choice of an

estimator of  $\tilde{z}_k$  is somewhat arbitrary under Deville's approach.

Demnati and Rao (2004) studied general estimators that can be expressed as smooth functions of the weights  $\mathbf{d}(s) = \{d_1(s), \dots, d_N(s)\}^T$ , say  $\hat{\theta} = f(\mathbf{d}(s))$ , and obtained a Taylor linearization variance estimator directly as  $\mathfrak{G}(z)$  with known linearized variables  $z_k = \partial f(b) / \partial b_k |_{b=\mathbf{d}(s)}$  without estimating  $\tilde{z}_k$  first and then replacing it by an estimator. For example, in the case of the ratio estimator their method automatically leads to  $z_k$  given above. This method can be applied to a variety of estimators including estimators of "census" logistic regression parameters based on calibration weights (Demnati and Rao 2004). Previous work on direct variance estimation includes Binder (1996).

When analyzing survey data, the population values  $y_k$ ,  $k = 1, \dots, N$ , are often assumed to be generated from a super-population model, and the user is often interested in making inferences on the model parameters. Let  $\theta_N$  be a "census" parameter, i.e., an estimator of a model parameter  $\theta$  when the population  $y_k$ -values are all known, and let  $\hat{\theta}$  be a design-unbiased estimator of  $\theta_N$ , the "census" parameter. Suppose that  $\hat{\theta}$  is design-model unbiased for  $\theta$ , i.e.,  $E_m E_p(\hat{\theta}) = \theta$ , where  $E_m$  and  $E_p$  respectively denote the expectations with respect to the design and the model. Then the total variance of  $\hat{\theta}$  is  $V(\hat{\theta}) = E_m E_p(\hat{\theta} - \theta)^2$  which can be decomposed as

$$V(\hat{\theta}) = E_m V_p(\hat{\theta}) + V_m(\theta_N), \quad (1.1)$$

where  $V_p(\hat{\theta}) = E_p(\hat{\theta} - \theta_N)^2$  is the design variance of  $\hat{\theta}$  and  $V_m(\theta_N)$  is the model variance of  $\theta_N$ . It follows from (1.1) that the total variance may be estimated using a design-based estimator of  $V_p(\hat{\theta})$  if the last term  $V_m(\theta_N)$  is negligible relative to  $E_m V_p(\hat{\theta})$ . In that case, the distinction between  $\theta_N$  and  $\theta$  can be ignored (Skinner, Holt and Smith 1989, page 14). On the other hand, it is necessary to estimate the total variance  $V(\hat{\theta})$  when the model variance  $V_m(\theta_N)$  is not negligible relative to  $E_m V_p(\hat{\theta})$ . This requires consideration of the joint design and model random processes. Molina, Smith and Sugden (2001) argued that the combined process of generation of the finite population and selection of the sample should be the basis for analytical inferences on model parameters. Rubin-Bleuer and Şchiopu-Kratina (2005) have provided a mathematical framework for joint model and design-based inference. However, a broadly applicable method is needed for the estimation of total variance. The main purpose of this paper is to provide such a method, by extending the Demnati-Rao approach for finite population parameters.

In Section 2, we consider the case of a scalar parameter  $\theta$  and present linearization variance estimators by expanding the Demnati and Rao (2004) approach. The

method is illustrated for the special case of a ratio estimator of a super-population mean  $\theta$ . Results of Section 2 are extended in Section 3 to estimators of a vector parameter  $\boldsymbol{\theta}$  obtained as solutions to weighted estimating equations, and the method is illustrated for the special case of parameters of a logistic regression model. Simulation results are also presented.

## 2. Scalar model parameter

### 2.1 Point estimators

Consider a finite population  $U$  of  $N$  elements, and let  $d_k(s) = a_k(s) / \pi_k$  be the design weights attached to the population element  $k$ , where  $a_k(s) = 1$  if element  $k$  is in the sample  $s$  and  $a_k(s) = 0$  otherwise, and  $\pi_k$  is the inclusion probability associated with  $k$ . We consider estimators  $\hat{\theta}$  of a scalar parameter  $\theta$  that can be expressed as functions of random variables under the design and the assumed model. In particular,  $\hat{\theta} = f(\mathbf{A}_d)$ , where  $\mathbf{A}_d$  is a  $(p+1) \times N$  matrix with columns  $\mathbf{d}_k = (d_k h_{1k}, d_k h_{2k}, \dots, d_k h_{(p+1)k})^T \equiv (d_{1k}, \dots, d_{(p+1)k})^T$  where  $d_k = d_k(s)$  is random under the design,  $h_{1k} = 1$ , and  $h_{ik}$  ( $i = 2, \dots, p+1$ ) are random under the model.

For example, consider the ratio model with fixed covariates  $x_k$ :

$$E_m(y_k) = \beta x_k, \quad V_m(y_k) = \sigma^2 x_k, \quad \text{Cov}_m(y_k, y_t) = 0, \\ k \neq t, \quad k, t = 1, \dots, N, \quad (2.1)$$

where  $E_m$ ,  $V_m$ , and  $\text{Cov}_m$  denote model expectation, model variance, and model covariance respectively and  $\sigma^2 > 0$ . Suppose that we are interested in estimating the super-population mean  $\theta = E_m(\bar{Y}) = N^{-1} \sum E_m(y_k) = \beta \bar{X}$  where  $\bar{Y}$  is the finite population mean of  $y$ . In this case, a ratio estimator of  $\theta$  is given by

$$\hat{\theta} = \bar{X}(\hat{Y}/\hat{X}) \equiv \bar{X}\hat{R}, \quad (2.2)$$

where  $\hat{Y} = \sum d_k(s)y_k$  and  $\hat{X} = \sum d_k(s)x_k$  are the design-unbiased estimators of the totals  $Y$  and  $X$ , and  $\bar{X}$  is the known population mean of  $x$ . We can write the ratio estimator (2.2) in the form  $\hat{\theta} = \bar{X}(\sum d_{2k}) / \sum d_{1k} x_k$ , where  $d_{1k} = d_k(s)$  and  $d_{2k} = d_k(s)y_k$ . This is a special case of  $f(\mathbf{A}_d)$  with  $p=1$  and  $h_{2k} = y_k$ .

Let  $E_p$  be the design expectation and  $E = E_m E_p$  be the total expectation. Then, we have  $E(d_{1k}) = E_m(1) = 1 \equiv \mu_{1k}$  and  $E(d_{ik}) = E_m(g_{ik}) \equiv \mu_{ik}$ ,  $i = 2, \dots, p+1$ , noting that  $E_p(d_k(s)) = 1$ . We assume that  $f(\mathbf{A}_\mu) = \theta$ , where  $\mathbf{A}_\mu$  is a  $(p+1) \times N$  matrix with columns  $\boldsymbol{\mu}_k = (\mu_{1k}, \mu_{2k}, \dots, \mu_{(p+1)k})^T$ . Hence,  $\hat{\theta}$  is asymptotically  $pm$ -unbiased for  $\theta$ . In the special case of the ratio estimator, we have  $f(\mathbf{A}_\mu) = \beta \bar{X} = \theta$ , noting that  $\mu_{1k} = 1$  and  $\mu_{2k} = \beta x_k$ .



## 2.2 Linearization variance estimator

We first derive an estimator of the total variance of a linear estimator  $\hat{U} = \sum \mathbf{u}_k^T \mathbf{d}_k$ , where  $\mathbf{u}_k$  is a vector of constants. The total variance of  $\hat{U}$  may be decomposed as

$$V(\hat{U}) = E_m V_p(\hat{U}) + V_m E_p(\hat{U}) \equiv I + II, \quad (2.3)$$

where  $V_p$  and  $V_m$  denote design variance and model variance respectively. A design-unbiased estimator of the component  $I$  of the total variance (2.3) is obtained by estimating the design variance  $V_p(\hat{U})$  for fixed  $\mathbf{h}_k = (h_{1k}, \dots, h_{(p+1)k})^T$ . Now, noting that  $\hat{U} = \sum b_k d_k(s)$  is the standard Narain-Horvitz-Thompson (NHT) estimator of the total  $U = \sum b_k$  when  $b_k = \mathbf{u}_k^T \mathbf{h}_k$  are fixed conditionally, we can use either the Sen-Yates-Grandy (SYG) variance estimator for fixed sample size designs or the Horvitz-Thompson (HT) variance estimator for arbitrary designs. The SYG estimator is given by

$$\begin{aligned} \text{est}(I) &= \mathfrak{g}_{\text{SYG}}(\hat{U}) \\ &= \sum \sum_{k < t} d_{kt}(s) \frac{(\pi_k \pi_t - \pi_{kt})}{\pi_k \pi_t} (b_k - b_t)^2, \end{aligned} \quad (2.4)$$

where  $d_{kt}(s) = \{a_k(s) a_t(s)\} / \pi_{kt}$  and  $\pi_{kt}$  is the inclusion probability for units  $k$  and  $t$  ( $k \neq t$ ). The HT variance estimator is given by

$$\text{est}(I) = \mathfrak{g}_{\text{HT}}(\hat{U}) = \sum \sum d_{kt}(s) \frac{(\pi_{kt} - \pi_k \pi_t)}{\pi_k \pi_t} b_k b_t, \quad (2.5)$$

where  $d_{kk}(s) = d_k(s)$ . For the special case of stratified random sampling (2.4) and (2.5) are identical.

Turning to the component  $II$  of the total variance (2.3), we have  $V_m E_p(\hat{U}) = V_m(\sum \mathbf{u}_k^T \mathbf{h}_k) = \sum \sum \mathbf{u}_k^T \text{Cov}_m(\mathbf{h}_k, \mathbf{h}_t) \mathbf{u}_t$  and a  $pm$ -unbiased estimator is therefore given by

$$\text{est}(II) = \sum \sum d_{kt}(s) \mathbf{u}_k^T \text{cov}_m(\mathbf{h}_k, \mathbf{h}_t) \mathbf{u}_t, \quad (2.6)$$

after replacing  $\text{Cov}_m(\mathbf{h}_k, \mathbf{h}_t)$  by an estimator  $\text{cov}_m(\mathbf{h}_k, \mathbf{h}_t)$ . The estimator of total variance (2.3) is now given by  $\text{est}(I) + \text{est}(II)$ . We denote it, in operator notation, as  $\mathfrak{g}(\mathbf{u})$ .

We now turn to the estimation of total variance of  $\hat{\theta}$ . Following Demnati and Rao (2004), a Taylor expansion of  $\hat{\theta} - \theta$  may be written as

$$\hat{\theta} - \theta \approx \sum \tilde{\mathbf{z}}_k^T (\mathbf{d}_k - \boldsymbol{\mu}_k) \quad (2.7)$$

where  $\tilde{\mathbf{z}}_k = \partial f(\mathbf{A}_b) / \partial \mathbf{b}_k |_{\mathbf{A}_b = \mathbf{A}_d}$  and  $\mathbf{A}_b$  is a  $(p+1) \times N$  matrix with  $k^{\text{th}}$  column  $\mathbf{b}_k$ , a vector of arbitrary real numbers. The approximation (2.7) is valid for any  $\hat{\theta}$  that can be expressed as a smooth function of estimated totals. Following Demnati and Rao (2004), a linearization estimator of the total variance is now given by

$$\mathfrak{g}_{\text{DR}}(\hat{\theta}) = \mathfrak{g}(\tilde{\mathbf{z}}), \quad (2.8)$$

which is obtained from  $\mathfrak{g}(\mathbf{u})$  by replacing  $\mathbf{u}_k$  by the “linearized variable”  $\tilde{\mathbf{z}}_k = \partial f(\mathbf{A}_b) / \partial \mathbf{b}_k |_{\mathbf{A}_b = \mathbf{A}_d}$ . A rigorous theoretical justification of (2.8) follows along the lines of Deville (1999).

## 2.3 Special case of ratio estimator

For the ratio estimator  $\hat{\theta} = \bar{X} \hat{R}$  of the model parameter  $\theta = \beta \bar{X}$ ,  $\tilde{\mathbf{z}}_k$  reduces to

$$\tilde{\mathbf{z}}_k = (\bar{X} / \hat{X})(-\hat{R} x_k, 1)^T = (z_{1k}, z_{2k})^T. \quad (2.9)$$

Further,  $b_k$  in (2.4) or (2.5) is replaced by

$$\begin{aligned} \mathbf{z}_k^T \mathbf{h}_k &= z_{1k} + z_{2k} y_k \\ &= (\bar{X} / \hat{X})(y_k - \hat{R} x_k) \equiv (\bar{X} / \hat{X}) e_k, \end{aligned}$$

using (2.9). Also, replacing  $\mathbf{u}_k$  by  $\tilde{\mathbf{z}}_k$  in (2.6) we get

$$\mathbf{z}_k^T \text{cov}_m(\mathbf{h}_k, \mathbf{h}_t) \mathbf{z}_t = z_{2k} z_{2t} \text{cov}_m(y_k, y_t).$$

Under the ratio model (2.1) with unspecified model variance  $V_m(y_k) = \sigma_k^2$ ,  $k = 1, \dots, N$ , we can estimate  $\sigma_k^2 = E_m(y_k - \beta x_k)^2$  by  $(y_k - \hat{R} x_k)^2$  and letting  $\text{cov}_m(y_k, y_t) = 0$ , for  $k \neq t$ .

We now study the special case of simple random sampling without replacement. In this case, both (2.4) and (2.5) reduce to

$$\text{est}(I) = \left( \frac{\bar{X}}{\bar{x}} \right)^2 \frac{1}{n} \left( 1 - \frac{n}{N} \right) s_e^2, \quad (2.10)$$

where  $s_e^2 = \sum a_k(s) e_k^2 / (n-1)$ , and (2.6) reduces to

$$\text{est}(II) = \left( \frac{\bar{X}}{\bar{x}} \right)^2 \frac{(n-1)}{nN} s_e^2. \quad (2.11)$$

Hence, using (2.10) and (2.11), the variance estimator (2.8) reduces to

$$\begin{aligned} \mathfrak{g}_{\text{DR}}(\hat{\theta}) &= \text{est}(I) + \text{est}(II) \\ &= \left( \frac{\bar{X}}{\bar{x}} \right)^2 \frac{1}{n} \frac{N-1}{N} s_e^2. \end{aligned} \quad (2.12)$$

It is interesting to note that the “ $g$ -weight”  $\bar{X} / \bar{x}$  appears automatically in  $\mathfrak{g}_{\text{DR}}(\hat{\theta})$ , given by (2.12), and that the finite population correction  $1 - n/N$  is absent in  $\mathfrak{g}_{\text{DR}}(\hat{\theta})$  unlike in  $\text{est}(I)$  given by (2.10).

In the customary approach to the estimation of total variance (see e.g., Korn and Graubard 1998)  $V(\hat{\theta})$  is first written as



$$\begin{aligned} V(\hat{\theta}) &= E_m V_p(\hat{\theta}) + V_m E_p(\hat{\theta}) \\ &\approx E_m V_p(\hat{\theta}) + V_m(\bar{Y}) \\ &= E_m V_p(\hat{\theta}) + N^{-2} \sum E_m (y_k - \beta x_k)^2, \quad (2.13) \end{aligned}$$

under the ratio model with unspecified  $\sigma_k^2$ ,  $k = 1, \dots, N$ . The first term  $E_m V_p(\hat{\theta})$  in (2.13) is then estimated by a design-consistent estimator of  $V_p(\hat{\theta})$ , typically by (2.10) without the  $g$ -factor  $(\bar{X}/\bar{x})^2$ . The second term is estimated by  $N^{-2} \sum d_k(s)(y_k - \hat{R}x_k)^2 = (nN)^{-1}(n-1)s_e^2$ . The sum of the two estimated terms then equals (2.12) without the  $g$ -factor. We denote this customary variance estimator by  $\mathfrak{G}_{\text{cus}}(\hat{\theta})$ . On the other hand, if (2.10) with the  $g$ -factor is used to estimate  $V_p(\hat{\theta})$ , the sum of this estimated term and the previous estimator of the second term leads to a “hybrid” variance estimator

$$\mathfrak{G}_{\text{mix}}(\hat{\theta}) = \text{est}(I) + (nN)^{-1}(n-1)s_e^2,$$

where the  $g$ -term is absent in the last term. It is clear from the above results that the choice of estimator of total variance under the customary approach is not unique, unlike under the proposed approach.

If the parameter of interest is  $\beta = \theta/\bar{X}$  instead of  $\theta$ , then  $\hat{\beta} = \hat{\theta}/\bar{X} = \hat{R}$  and  $\mathfrak{G}_{\text{DR}}(\hat{\beta})$  under simple random sampling is give by

$$\mathfrak{G}_{\text{DR}}(\hat{\beta}) = \bar{X}^{-2} \mathfrak{G}_{\text{DR}}(\hat{\theta}) = \bar{x}^{-2} \frac{1}{n} \frac{N-1}{N} s_e^2. \quad (2.14)$$

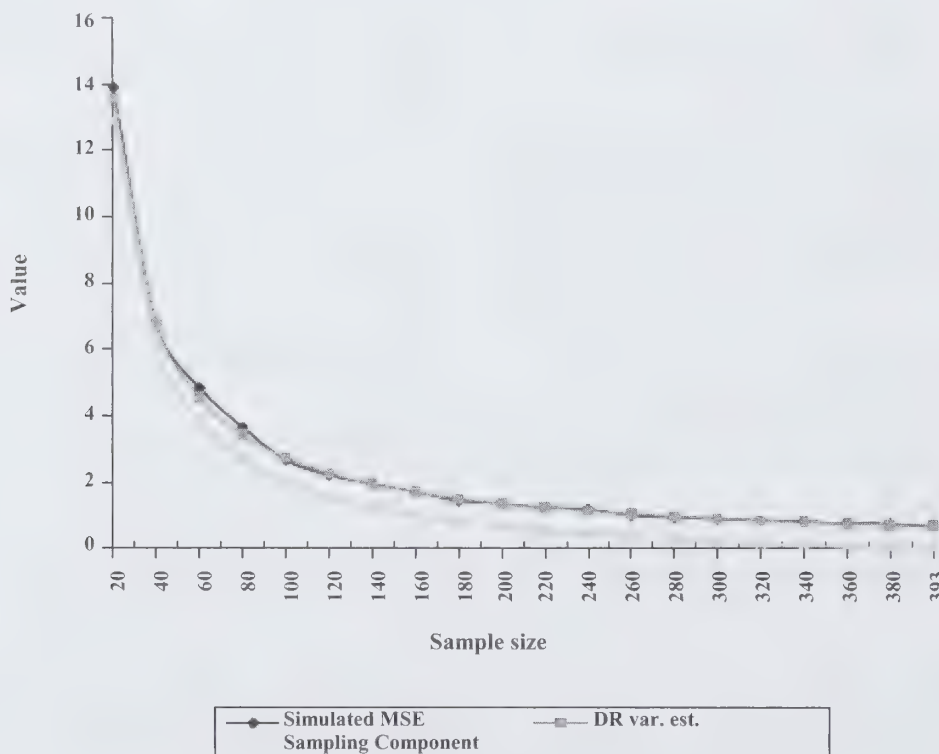
The customary approach leads to the same variance estimator, (2.14).

## 2.4 Simulation study

We conducted a small simulation study to examine the performances of different variance estimators, both unconditionally and conditionally on  $\hat{X}$ . We first generated  $R = 2,000$  finite populations  $\{y_1, \dots, y_N\}$  each of size  $N = 393$ , from the ratio model

$$y_k = 2x_k + x_k^{1/2}\varepsilon_k, \quad (2.15)$$

with independent values  $\varepsilon_k$  generated from  $N(0, 1)$ , where the fixed  $x_k$  are the “number of beds” for the Hospitals population studied in Valliant, Dorfman and Royall (2000, page 424-427). One simple random sample of specified size  $n$  is drawn from each generated population. Our parameter of interest is  $\theta = \beta\bar{X}$ , where  $\beta = 2$ .



**Figure 1** Averages of variance estimates for selected sample sizes compared to estimated MSE of the ratio estimator.  $\mathfrak{G}_{\text{DR}}$  = DR var. est.,  $\mathfrak{G}_s$  = Sampling component: ratio model

Simulated total MSE of the ratio estimator  $\hat{\theta} = \bar{X}(\bar{y}/\bar{x})$  is calculated as  $M(\hat{\theta}) = R^{-1} \sum_{r=1}^{2,000} (\hat{\theta}_r - \theta)^2$ , where  $\hat{\theta}_r$  is the value of  $\hat{\theta}$  for the  $r^{\text{th}}$  simulated sample and  $(\bar{y}, \bar{x})$  are the sample means. We calculated the total variance estimate  $\vartheta_{\text{DR}}(\hat{\theta})$ , and its components  $\vartheta_s = \text{est}(I)$  and  $\vartheta_m = \text{est}(II)$  from each simulated sample  $r$  and their averages  $\bar{\vartheta}_{\text{DR}}$ ,  $\bar{\vartheta}_s$ , and  $\bar{\vartheta}_m$  over  $r$ . Figure 1 gives a plot of the average of variance estimates,  $\bar{\vartheta}_{\text{DR}}$  and  $\bar{\vartheta}_s$ , and the simulated total MSE for  $n = 20, 40, \dots, 380, 393$ . In the case of  $n = N$ ,  $\bar{\vartheta}_s = 0$ . It is seen from Figure 1, that  $\vartheta_{\text{DR}}$  is approximately unbiased, whereas  $\vartheta_s$  leads to severe underestimation as the sample size,  $n$ , increases.

We also examined the conditional performance of the variance estimators under simple random sampling given  $\bar{x}$ , by conducting another simulation study for inference on  $\theta$ , using model (2.15). The study is similar to the study of Royall and Cumberland (1981) for inference on the finite population mean  $\theta_N = \bar{Y}$  from a fixed population  $\{y_1, \dots, y_N\}$ . We generated  $R = 20,000$  finite populations  $\{y_1, \dots, y_N\}$ , each of size  $N = 393$  from (2.15) using the number of beds as  $x_k$ , and from each population we then selected one simple random sample of size  $n = 100$ . We arranged the 20,000 samples in ascending order of  $\bar{x}$ -values and then grouped them into 20 groups each of size 1,000 such that the first group,  $G_1$ , contained 1,000 samples with the smallest  $\bar{x}$ -values, the next group,  $G_2$ , contained the next 1,000 smallest  $\bar{x}$ -values, and so on to get  $G_1, \dots, G_{20}$ . For each of the 20 groups so formed, we calculated the average values of the ratio estimates  $\hat{\theta} = \bar{X}(\bar{y}/\bar{x})$  and the mean estimates  $\bar{y}$ , and the resulting

conditional relative bias (CRB) in estimating  $\theta = 2\bar{X}$ ; see Figure 2. It is clear from Figure 2 that  $\bar{y}$  is conditionally biased unlike  $\hat{\theta}$ : negative CRB (-14%) for  $G_1$  increasing to positive CRB (+14%) for  $G_{20}$ . Note that both  $\bar{y}$  and  $\hat{\theta}$  are unconditionally unbiased for  $\theta$ . The conditional bias of  $\hat{\theta}$  and  $\bar{y}$  in estimating the model parameter  $\theta$  is similar to the conditional bias in estimating the "census" parameter  $\theta_N = \bar{Y}$ , as observed by Royall and Cumberland (1981).

We also calculated the conditional MSE of  $\hat{\theta}$  and the associated CRB of the variance estimators  $\vartheta_{\text{DR}}$ ,  $\vartheta_{\text{cus}}$  and  $\vartheta_{\text{mix}}$  based on the average values of  $\vartheta_{\text{DR}}$ ,  $\vartheta_{\text{cus}}$  and  $\vartheta_{\text{mix}}$  in each group; see Figure 3. It is evident from Figure 3 that CRB of  $\vartheta_{\text{cus}}$  ranges from -28% to 20% across the groups whereas  $\vartheta_{\text{DR}}$  exhibits no such trend and its CRB is less than 5% in absolute value except for  $G_6$  and  $G_{20}$ . Also, the CRB of  $\vartheta_{\text{mix}}$  is largely negative and below that of  $\vartheta_{\text{DR}}$  for the first half of the groups and above for the second half, but  $\vartheta_{\text{mix}}$  exhibits no visible trends unlike  $\vartheta_{\text{cus}}$ .

Figure 4 reports the conditional coverage rates (CCR) of normal theory confidence intervals based on  $\vartheta_{\text{DR}}$ ,  $\vartheta_{\text{cus}}$ ,  $\vartheta_{\text{mix}}$  and  $\vartheta_s$  (ignoring the component  $\vartheta_m$ ) for nominal level of 95%. As expected, the use of  $\vartheta_s$  leads to severe undercoverage because the sampling fraction,  $100/393$ , is significant. On the other hand, CCR associated with  $\vartheta_{\text{DR}}$  is closer to nominal level across groups, while  $\vartheta_{\text{cus}}$  exhibits a trend across groups with CCR ranging from 91% to 97%. Further, CCR associated with  $\vartheta_{\text{mix}}$  is slightly below that of  $\vartheta_{\text{DR}}$  for the first half of the groups but  $\vartheta_{\text{mix}}$  and  $\vartheta_{\text{DR}}$  perform similarly.

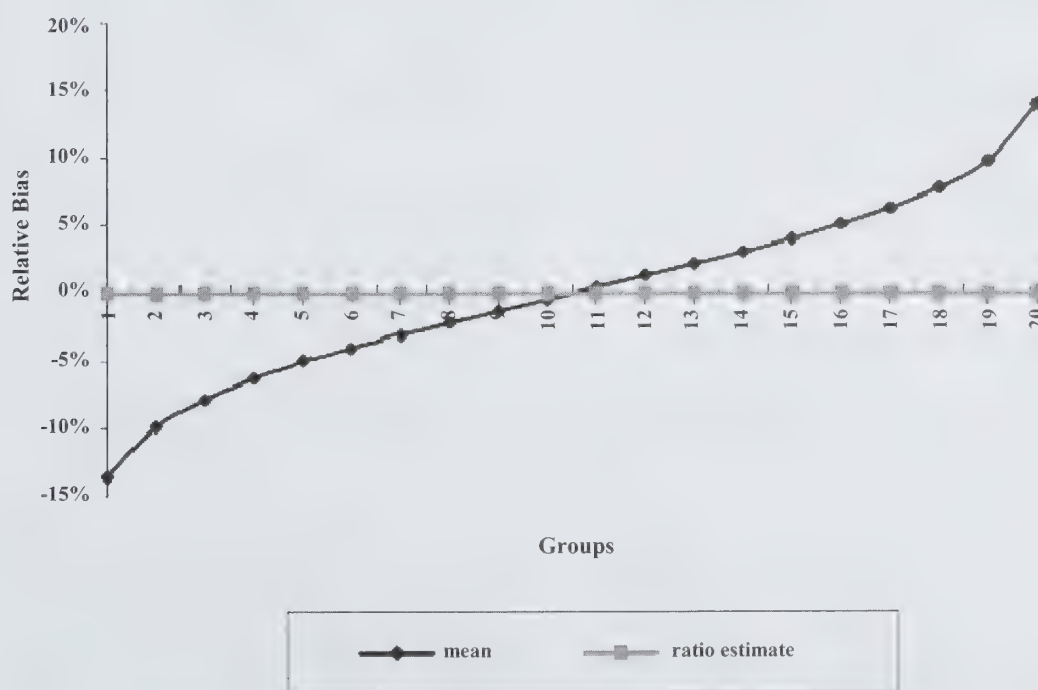


Figure 2 Conditional relative bias of the expansion and ratio estimators: ratio model

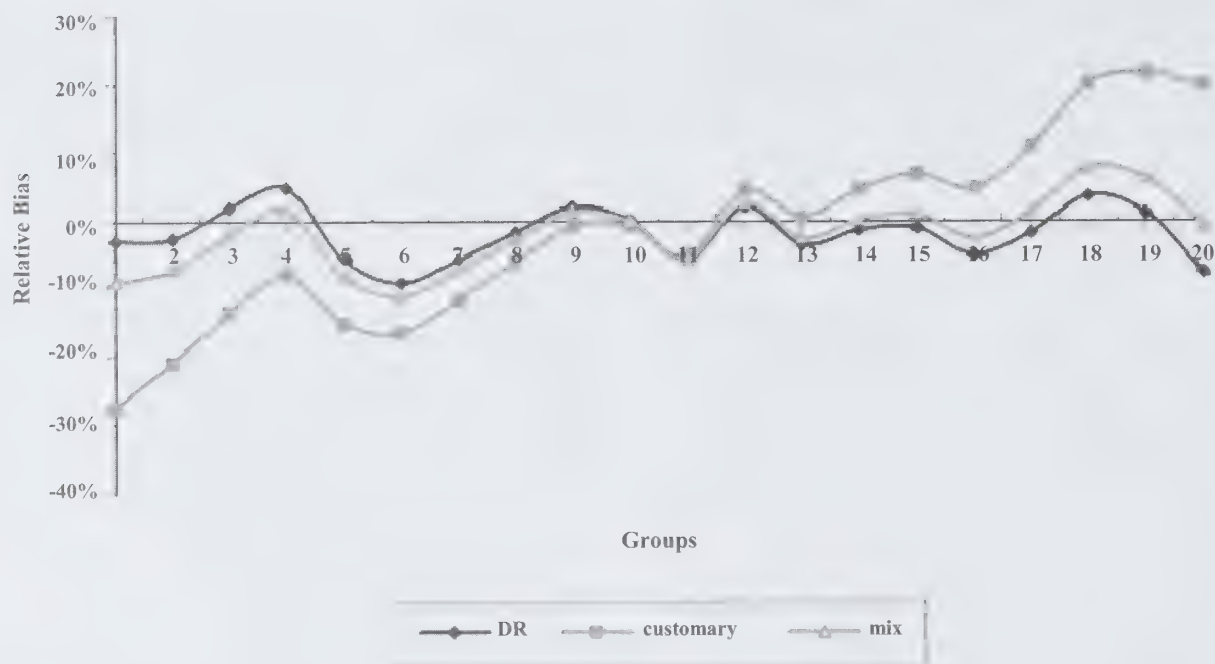


Figure 3 Conditional relative bias of variance estimators  $\vartheta_{DR}$ ,  $\vartheta_{cus}$  and  $\vartheta_{mix}$ : ratio model

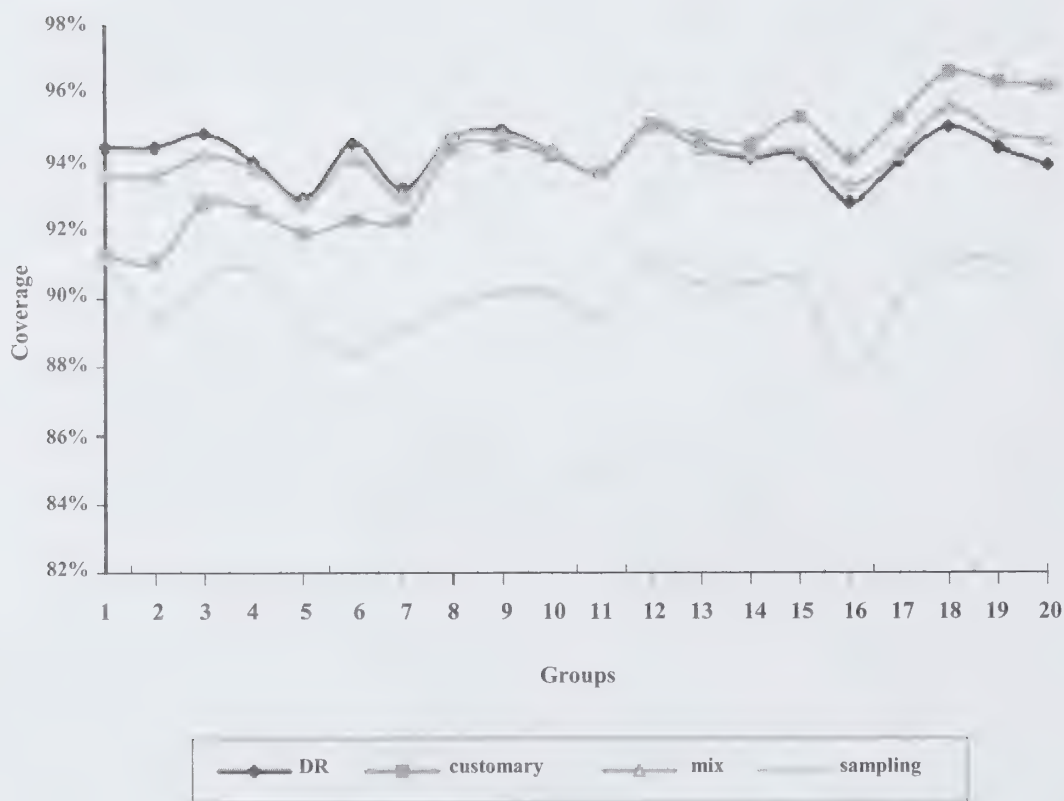


Figure 4 Conditional coverage rates of normal theory confidence intervals based on  $\vartheta_{DR}$ ,  $\vartheta_{cus}$ ,  $\vartheta_{mix}$  and  $\vartheta_s$  for nominal level of 95%: ratio model



### 3. Calibration weighted estimating equations

#### 3.1 Estimators of model parameters

Suppose that the super-population model on the responses  $y_k$  is specified by a generalized linear model (McCullagh and Nelder 1989) with mean  $E_m(y_k) = \mu_k(\boldsymbol{\theta}) = h(\mathbf{x}_k^T \boldsymbol{\theta})$ , where  $\mathbf{x}_k$  is a  $p \times 1$  vector of explanatory variables,  $\boldsymbol{\theta}$  is the  $p$ -vector of model parameters and  $h(\cdot)$  is a “link” function. For example,  $h(a) = a$  gives a linear regression model and  $h(a) = e^a / (1 + e^a)$  gives a logistic regression model for binary responses  $y_k$ .

We define census estimating equations (CEE), based on estimating functions  $I_k(\boldsymbol{\theta})$ , as  $I(\boldsymbol{\theta}) = \sum I_k(\boldsymbol{\theta}) = \mathbf{0}$  with  $E_m I_k(\boldsymbol{\theta}) = \mathbf{0}$ , and the solution to CEE gives the census parameter vector  $\boldsymbol{\theta}_N$ . For example,  $I_k(\boldsymbol{\theta}) = \mathbf{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$  for linear and logistic regression models. We use generalized regression (GREG) weights  $w_k(s) = d_k(s)g_k(d(s))$ , where the “g-weights” are given by

$$g_k(d(s)) = 1 + (\mathbf{T} - \hat{\mathbf{T}})^T \left[ \sum d_k(s) c_k \mathbf{t}_k \mathbf{t}_k^T \right]^{-1} c_k \mathbf{t}_k,$$

for specified  $c_k$ , where  $\hat{\mathbf{T}} = \sum d_k(s) \mathbf{t}_k$  is the HT estimator of the known total  $\mathbf{T}$  of a  $q \times 1$  vector of calibration variables  $\mathbf{t}_k$  and  $d(s)$  is the  $N \times 1$  vector of the weights  $d_k(s)$ . The GREG weights,  $w_k(s)$ , have the calibration property  $\sum w_k(s) \mathbf{t}_k = \mathbf{T}$  and lead to efficient estimators  $\tilde{Y} = \sum w_k(s) y_k$  of totals  $Y = \sum y_k$ , when  $y_k$  and  $\mathbf{t}_k$  are linearly related (Särndal, Swensson and Wretman 1989, chapter 6).

We use the calibration weights,  $w_k(s)$ , to estimate the CEE. The calibration weighted estimating equations are given by

$$\tilde{I}(\boldsymbol{\theta}) = \sum w_k(s) I_k(\boldsymbol{\theta}) = \sum d_k(s) g_k(d(s)) I_k(\boldsymbol{\theta}) = \mathbf{0}. \quad (3.1)$$

The solution to (3.1), obtained by the Newton-Raphson-type iterative method, gives the calibration-weighted estimator  $\tilde{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$ , and  $\tilde{\boldsymbol{\theta}}$  is approximately design-model unbiased for  $\boldsymbol{\theta}$ , i.e.,  $E(\tilde{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}$ . It follows from (3.1) that  $\tilde{\boldsymbol{\theta}}$  is of the form  $\mathbf{f}(\mathbf{A}_d)$  with  $\mathbf{d}_k = (d_k(s), d_k(s) \mathbf{I}_k^T(\boldsymbol{\theta}))^T$ , where  $\mathbf{f}(\mathbf{A}_d)$  is a  $p \times 1$  vector and  $\mathbf{A}_d$  is a  $(p+1) \times N$  matrix with  $k^{\text{th}}$  column  $\mathbf{d}_k$ . Here we have  $h_k = 1$  and  $(h_{2k}, \dots, h_{(p+1)k}) = \mathbf{I}_k(\boldsymbol{\theta})$ .

#### 3.2 Linearized variance estimators

We first extend the result on variance estimation for the scalar case  $\hat{U} = \sum \mathbf{b}_k^T \mathbf{d}_k$  (Section 2.2) to the vector case  $\hat{U} = \sum \mathbf{U}_k \mathbf{d}_k = \sum \mathbf{b}_k^T \mathbf{d}_k(s)$ , where  $\mathbf{b}_k = \mathbf{U}_k \mathbf{h}_k$  is a  $p$ -vector and  $\mathbf{U}_k$  is a  $p \times (p+1)$  matrix with rows  $\mathbf{u}_{jk}^T$ ,  $j=1, \dots, p$ . In this case, the SYG variance estimator (2.4) is changed to

$$\text{est}(I) = \mathfrak{G}_{\text{SYG}}(\hat{U})$$

$$= \sum \sum_{k < t} d_{kt}(s) \frac{(\pi_k \pi_t - \pi_{kt})}{\pi_k \pi_t} (\mathbf{b}_k - \mathbf{b}_t)(\mathbf{b}_k - \mathbf{b}_t)^T. \quad (3.2)$$

Similarly, the H-T variance estimator (2.5) is changed to

$$\text{est}(I) = \mathfrak{G}_{\text{HT}}(\hat{U}) = \sum \sum d_{kt}(s) \frac{(\pi_{kt} - \pi_k \pi_t)}{\pi_k \pi_t} \mathbf{b}_k \mathbf{b}_t^T. \quad (3.3)$$

Turning to the component  $II$  of the total variance of  $\hat{U}$ , (2.6) is changed to

$$\text{est}(II) = \sum \sum d_{kt}(s) \mathbf{U}_k \text{cov}_m(\mathbf{h}_k, \mathbf{h}_t) \mathbf{U}_t^T. \quad (3.4)$$

The total variance of  $\hat{U}$  is estimated by the sum of (3.2) and (3.4) for fixed sample size designs or by the sum of (3.3) and (3.4) for arbitrary designs.

A linearization variance estimator of the total variance of  $\tilde{\boldsymbol{\theta}}$  is obtained from the estimated total variance estimator of  $\hat{U}$  by replacing  $\mathbf{U}_k$  by the linearized variable  $\mathbf{Z}_k = \partial \mathbf{f}(\mathbf{A}_d) / \partial \mathbf{b}_k |_{\mathbf{A}_d = \mathbf{A}_d}$ . Following the implicit differentiation method of Demnati and Rao (2004),  $\mathbf{Z}_k$  reduces to

$$\mathbf{Z}_k = [\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}})]^{-1} g_k(d(s)) (-\hat{\mathbf{B}}_I^T \mathbf{t}_k, \mathbf{I}_p),$$

with

$$\hat{\mathbf{B}}_I = \left[ \sum d_k(s) c_k \mathbf{t}_k \mathbf{t}_k^T \right]^{-1} \sum d_k(s) c_k \mathbf{t}_k \mathbf{t}_k^T (\tilde{\boldsymbol{\theta}}),$$

$$\tilde{\mathbf{J}}(\boldsymbol{\theta}) = -\sum d_k(s) g_k(d(s)) (\partial I_k(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T),$$

and  $\mathbf{I}_p$  is the  $p \times p$  identity matrix.

After some simplification, the first component  $\text{est}(I)$  is given by (3.2) or (3.3) with  $\mathbf{b}_k$  changed to

$$\mathbf{Z}_k \mathbf{h}_k = [\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}})]^{-1} \mathbf{e}_k(\tilde{\boldsymbol{\theta}}) g_k(d(s)), \quad (3.5)$$

where

$$\mathbf{e}_k(\tilde{\boldsymbol{\theta}}) = \mathbf{I}_k(\tilde{\boldsymbol{\theta}}) - \hat{\mathbf{B}}_I^T \mathbf{t}_k.$$

Similarly, the second component  $\text{est}(II)$  simplifies to

$$\text{est}(II) =$$

$$[\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}})]^{-1} \sum d_k(s) g_k^2(d(s)) \mathbf{I}_k(\tilde{\boldsymbol{\theta}}) \mathbf{I}_k^T(\tilde{\boldsymbol{\theta}}) [\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}})]^{-1}, \quad (3.6)$$

if  $\text{Cov}_m[\mathbf{I}_k(\tilde{\boldsymbol{\theta}}) \mathbf{I}_t^T(\tilde{\boldsymbol{\theta}})] = \mathbf{0}$  for  $k \neq t$ .

The total variance estimator of  $\tilde{\boldsymbol{\theta}}$  is now estimated by

$$\mathfrak{G}_{\text{DR}}(\tilde{\boldsymbol{\theta}}) = \text{est}(I) + \text{est}(II). \quad (3.7)$$

This variance estimator of  $\tilde{\boldsymbol{\theta}}$  automatically takes account of the g-weights as in Section 2.

A customary variance estimator of  $\tilde{\boldsymbol{\theta}}$ ,  $\mathfrak{G}_{\text{cus}}(\tilde{\boldsymbol{\theta}})$ , is obtained from (3.7) by ignoring the g-weights in (3.5) and (3.6). Similarly, a hybrid variance estimator,  $\mathfrak{G}_{\text{mix}}(\tilde{\boldsymbol{\theta}})$ , is

obtained from (3.7) by retaining the  $g$ -weights in  $\text{est}(I)$  and ignoring them in  $\text{est}(II)$ .

### 3.3 Simulation study

We conducted a simulation study to compare the relative performances of the three variance estimators  $\vartheta_{\text{DR}}$ ,  $\vartheta_{\text{cus}}$ , and  $\vartheta_{\text{mix}}$ , for the special case of a logistic regression model:

$$E_m(y_k) = \mu_k(\boldsymbol{\theta}) = \exp(\mathbf{x}_k^T \boldsymbol{\theta}) / \{1 + \exp(\mathbf{x}_k^T \boldsymbol{\theta})\} \quad (3.8)$$

$$V_m(y_k) = \mu_k(\boldsymbol{\theta})(1 - \mu_k(\boldsymbol{\theta})), \text{Cov}_m(y_k, y_t) = 0, k \neq t.$$

In this case, we have  $\mathbf{l}_k(\boldsymbol{\theta}) = \mathbf{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$ , and

$$\tilde{\mathbf{J}}(\boldsymbol{\theta}) = \sum d_k(s) g_k(d(s)) \mathbf{x}_k \mathbf{x}_k^T \mu_k(\boldsymbol{\theta})(1 - \mu_k(\boldsymbol{\theta})).$$

For the simulation study, we set  $\mathbf{x}_k = (1, x_k)^T$ , where the  $x_k$  denote the number of beds for the Hospitals population of size  $N = 393$  studied in Section 2.2. We implemented post-stratification by dividing the population into two classes with  $N_1 = 171$  hospitals  $k$  having  $x_k < 350$  in class 1 and  $N_2 = 122$  hospitals  $k$  with  $x_k \geq 350$  in class 2. Here,  $g_k(d(s)) = N_h / \hat{N}_h$ ,  $h = 1, 2$ , if  $k$  belongs to class  $h$ , where  $\hat{N}_h = \sum d_k(s) t_{hk}$  is the design-weight estimator of  $N_h$ , and  $\mathbf{t}_k = (t_{1k}, t_{2k})^T$  is the vector of class indicator variables  $t_{hk}$ .

We generated  $R = 40,000$  finite populations  $\{y_1, \dots, y_N\}$ , each of size  $N = 393$ , assuming the logistic regression model (3.8) with  $\boldsymbol{\theta} = (\theta_0, \theta_1)^T = (-1, 0.005)^T$ . The parameter of interest is  $\theta_1 = 0.005$ . From each generated population, we selected one simple random sample of size  $n = 150$ , and then obtained the calibration-weighted estimated  $\tilde{\theta}_1$  and associated variance estimators  $\text{est}(I) = \vartheta_s(\tilde{\theta}_1)$ ,  $\vartheta_{\text{DR}}(\tilde{\theta}_1)$ ,  $\vartheta_{\text{cus}}(\tilde{\theta}_1)$  and  $\vartheta_{\text{mix}}(\tilde{\theta}_1)$  from each sample  $r$ . We obtained the averages of the estimates and the variance estimates as  $\text{av}(\hat{\theta}_1) \approx 0.00514$ ,  $\text{av}(\vartheta_{\text{DR}}) \approx 0.0989$ ,

$\text{av}(\vartheta_{\text{cus}}) \approx 0.0987$ ,  $\text{av}(\vartheta_{\text{mix}}) \approx 0.0988$ , and  $\text{av}(\vartheta_s) \approx 0.0613$ . Also, the estimated total MSE of  $\hat{\theta}_1$  is equal to 0.0998. Hence, unconditionally the estimator  $\tilde{\theta}_1$  is approximately unbiased for  $\theta_1$ , and the bias of the three variance estimators  $\vartheta_{\text{DR}}$ ,  $\vartheta_{\text{cus}}$  and  $\vartheta_{\text{mix}}$  is negligible. On the other hand ignoring the second component and using only the first component,  $\text{est}(I) = \vartheta_s(\tilde{\theta}_1)$ , leads to severe underestimation, as expected.

We also examined the conditional performances of the three variance estimators along the line of Section 2.2. We arranged the 40,000 samples in ascending order of the sample size,  $n_i$ , in class 1, and then grouped the samples into twenty groups, each of size 2,000, such that the first group,  $G_1$ , contained the 2,000 samples with the smallest  $n_1$ -values, the second group,  $G_2$ , contained the 2,000 samples with the next smallest  $n_1$ -values, and so on to get twenty groups,  $G_1, \dots, G_{20}$ .

We calculated the conditional MSE of  $\tilde{\theta}_1$  and the associated conditional relative bias (CRB) of the variance estimators  $\vartheta_{\text{DR}}$ ,  $\vartheta_{\text{cus}}$  and  $\vartheta_{\text{mix}}$  based on the average values of  $\vartheta_{\text{DR}}$ ,  $\vartheta_{\text{cus}}$  and  $\vartheta_{\text{mix}}$  in each group; see Figure 5. We can see from Figure 5 that CRB of  $\vartheta_{\text{cus}}$  ranges from 20% to -20% across the groups, whereas  $\vartheta_{\text{DR}}$  exhibits no such trend and its CRB is less than 5% in absolute value except for two groups. Also, the CRB of  $\vartheta_{\text{mix}}$  exhibits a trend but less pronounced than  $\vartheta_{\text{cus}}$ . Figure 6 reports the conditional coverage rates (CCR) of normal theory intervals based on  $\vartheta_{\text{DR}}$ ,  $\vartheta_{\text{cus}}$  and  $\vartheta_{\text{mix}}$  for nominal level of 95%. We can see from Figure 6 that  $\vartheta_{\text{cus}}$  exhibits a trend across groups with CCR ranging from 97% to 92%, whereas CCR associated with  $\vartheta_{\text{DR}}$  is close to the nominal level across groups. Further, CCR associated with  $\vartheta_{\text{mix}}$  is slightly above that of  $\vartheta_{\text{DR}}$  for the first half of the groups and slightly below for the remaining groups.

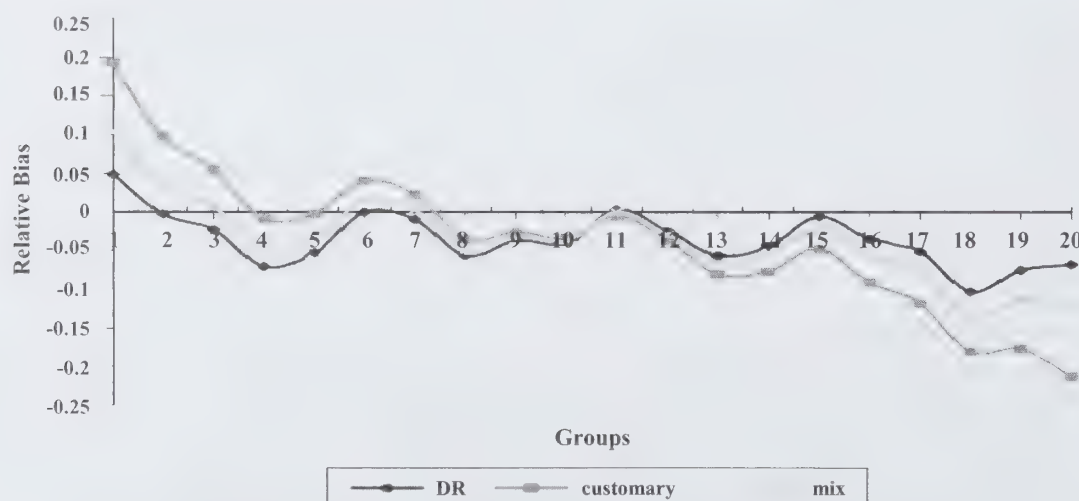


Figure 5 Conditional relative bias of variance estimators: logistic regression

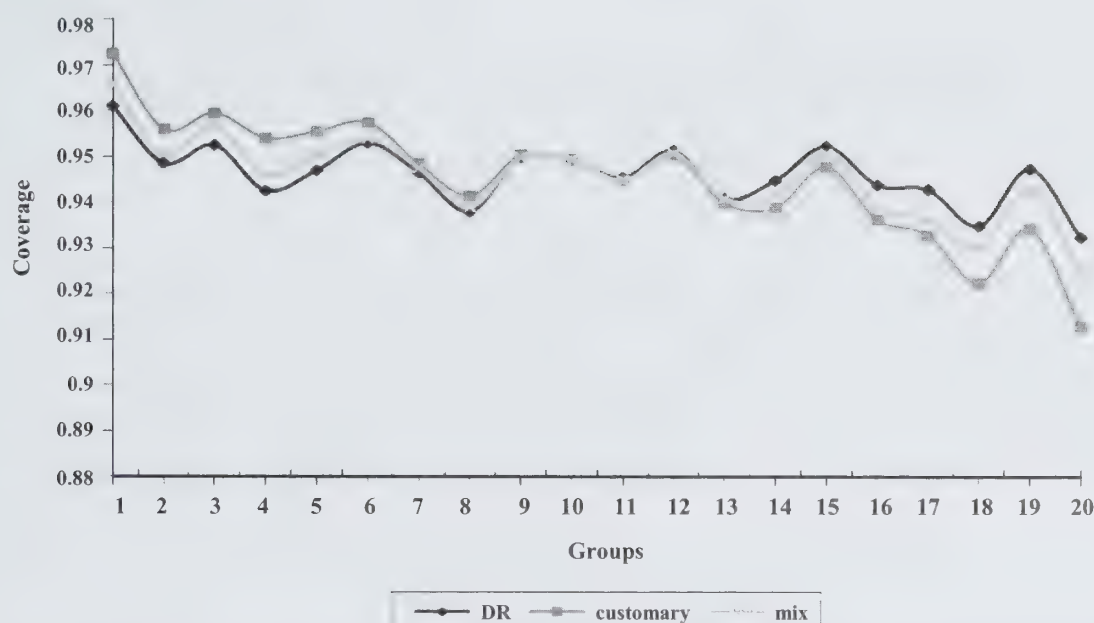


Figure 6 Conditional coverage rates of normal theory confidence intervals for nominal level of 95%: logistic regression

### Concluding remarks

We have studied the estimation of total variance of estimators of model parameters under an assumed super-population model. Our approach leads directly to a linearization variance estimator which is shown to perform well under a conditional framework when calibration weights are used for estimation. We are currently investigating extensions of our method to estimation of total variance under imputation for item nonresponse and integration of two independent surveys.

### Acknowledgements

We thank two referees for constructive comments and suggestions. J.N.K. Rao's work was partially supported by a grant from Natural Sciences and Engineering Research Council of Canada.

### References

- Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Binder, D. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.
- Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30, 17-34.
- Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.
- Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*, New York: John Wiley & Sons, Inc.
- McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, 2<sup>nd</sup> Ed. Chapman & Hall, London.
- Molina, E.A., Smith, T.M.F. and Sugden, R.A. (2001). Modeling overdispersion for complex survey data. *International Statistical Review*, 69, 373-384.
- Royall, R.M., and Cumberland, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.
- Rubin-Bleuer, S., and Şchiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, 33, 2789-2810.
- Särdal, C.-E., Swensson, B. and Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*, New York: John Wiley & Sons, Inc.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite population sampling and inference: A prediction approach*, New York: John Wiley & Sons, Inc.





# Statistical foundations of cell-phone surveys

Kirk M. Wolter, Phil Smith and Stephen J. Blumberg<sup>1</sup>

## Abstract

The size of the cell-phone-only population in the USA has increased rapidly in recent years and, correspondingly, researchers have begun to experiment with sampling and interviewing of cell-phone subscribers. We discuss statistical issues involved in the sampling design and estimation phases of cell-phone studies. This work is presented primarily in the context of a nonoverlapping dual-frame survey in which one frame and sample are employed for the landline population and a second frame and sample are employed for the cell-phone-only population. Additional considerations necessary for overlapping dual-frame surveys (where the cell-phone frame and sample include some of the landline population) are also discussed. We illustrate the methods using the design of the National Immunization Survey (NIS), which monitors the vaccination rates of children age 19-35 months and teens age 13-17 years. The NIS is a nationwide telephone survey, followed by a provider record check, conducted by the Centers for Disease Control and Prevention.

Key Words: Cell-phone study; Random digit dialing; Dual-frame survey; Network sampling; Indirect sampling; Linking rules; Weighting of survey data; National Immunization Survey.

## 1. Introduction

The number of persons with cell phones in the USA has increased rapidly in recent years, and the percent of adults living in households with cell phones is expected to soon exceed the percent living in households with landlines (CTIA 2008; Blumberg and Luke 2008; Arthur 2007; Ehlen and Ehlen 2007). Correspondingly, survey researchers have begun to experiment with the sampling and interviewing of cell-phone subscribers (Lavrakas, Shuttles, Steeh and Fienberg 2007). This article is about the issues of statistical design and estimation that arise in cell-phone surveys. It emphasizes theoretically rigorous but practical solutions to the emergent problems survey researchers are facing in cell-phone surveys today.

Standard telephone surveys driven by random-digit-dialing (RDD) sampling only cover the population of households that have at least one working landline telephone actually used for voice communications. In an RDD survey, one assumes that the landline telephone is a household appliance and that all persons in the population are attached to one and only one household. Thus, one can sample people indirectly by sampling their telephone numbers and proceed from there to use reasonably standard and well-known methods of estimation.

The cell-phone survey brings a paradigm shift and new challenges. Most people think of the cell phone as a personal appliance, not a household device. Some people do share a cell phone, including 10-20 percent of cell-phone-only adults (Carley-Baxter, Peytchev and Lynberg 2008), but many do not, and thus it cannot be assumed that all residents of a household can be reached through the same

cell-phone line. Some residents of a household can be reached through more than one cell-phone line. Some residents can be reached only by a cell-phone line while others can be reached through both cell and landline telephones. Thus, in the cell-phone survey, the household may no longer provide the same unifying organization that it does in standard telephone surveys.

To address the growing risk of bias (due to under-coverage) in telephone surveys, one can consider dual-frame telephone survey designs that include both an RDD sample of landline telephones and a sample of cell-phone lines. The telephone numbers on the two sampling frames are non-overlapping, but the corresponding people and households that may be the objects of the survey are partially overlapping.

A rigorous theory of estimation for such telephone survey designs has been lacking, although some initial descriptions of weighting have been advanced by Brick, Dipko, Presser, Tucker and Yuan (2006), Brick, Edwards and Lee (2007), and Frankel, Battaglia, Link and Mokdad (2007). In this article, we provide a general theory of unbiased estimation for population totals in the context of dual-frame telephone survey designs and derive the corresponding survey weights. We show what information must be collected in the survey itself to enable the calculation of the sampling weights.

To introduce ideas, we let  $A$  signify the portion of the overall population of interest accessible through the landline sampling frame, let  $B$  denote the portion accessible through the cell-phone sampling frame, and let  $C$  denote the portion not accessible through either frame (the *phoneless population* and other relatively small components of the

1. Kirk M. Wolter, NORC and the University of Chicago. E-mail: wolter-kirk@norc.org; Phil Smith, National Center for Immunization and Respiratory Diseases; Stephen J. Blumberg, National Center for Health Statistics.

total population). We let  $a$  be the subpopulation in  $A$  not accessible through cell-phone lines (the *landline-only population*), let  $b$  be the subpopulation in  $B$  not accessible through landlines (the *cell-phone-only population*), and let  $ab$  be the subpopulation accessible through both landlines and cell-phone lines (the *mixed population*). We will sharpen this notation in succeeding sections.

Whether or not a unit in the population of interest is accessible through landlines or cell-phone lines is itself a complex matter. Throughout this article, when we say that a unit is accessible through landlines, we shall mean that there is both physical access to one or more landlines (usually residential landlines only) and a respondent would actually answer the landline if it rang for voice communications. Many adults today maintain a landline telephone strictly for computer communications and utilize a cell phone for all voice communications. By our definition, such adults are not considered to have landline access and instead are considered to be in the cell-phone-only population. Similarly, when we say that a unit is accessible through cell-phone lines, we shall mean that there is both physical access to a cell phone and intent to answer the cell phone if it rang. All other units in the population of interest that are not accessible through either landlines or cell-phone lines are considered phoneless. Current evidence suggests, although no one knows for sure, that about 20 to 30 percent of adults are domain  $b$ , 5 to 10 percent are in domain  $C$ , and the balance are spread across domains  $a$  and  $ab$ .

What we know so far from the cell-phone surveys we and others have conducted is that the data collection is relatively expensive, with average-interviewer-hours-per-completed case running around three times the average for standard RDD surveys. The higher cost is brought, in part, by the legal requirement (in the US, the Telephone Consumer Protection Act) of manually dialing the selected cell-phones. Response rates are somewhat lower than those achieved in RDD surveys. Interview length may be problematic, with some respondents less willing to submit to a lengthy interview by cell phone than by landline phone. Privacy issues may constrain the cell-phone interview, if the respondent is not in a private place at the time of the interview. The cell-phone user's propensity to respond may vary monotonically with his or her level of use of the cell phone, with the heavy user more willing to answer the phone than the lighter or occasional user. Most breakoffs occur during the opening seconds of the interview attempt. Because cell-phone surveys are relatively new, people are not used to being called and the interviewer has mere seconds to sell the survey. On the other hand, we find many cell-phone respondents to be quite cooperative once their attention has been held through the survey's introductory script.

Due to all of these circumstances in the environment, we currently view the cell-phone sample as a relatively small supplementary sample, with the main sample continuing to be a larger RDD sample of landlines. The cell-phone sample is intended to round out the coverage of the population of interest. In the future, as the environment matures and if costs come down, it may be possible to shift towards a more balanced approach with similarly sized landline and cell-phone samples, or even to a state where the cell-phone sample begins to dominate and the landline sample is used as a supplement to round out coverage.

In Section 2, we introduce the topic of *networks* of *sampling units*, *reporting units*, and *estimation units* and show how cell-phone surveys equate to a sampling of networks. Section 3 introduces various key concepts that will be needed as we discuss survey estimation, among them being the idea of a *link* (or edge) between the *nodes* (or vertices) in the network. Section 4 describes the duality that exists between the populations corresponding to the different types of nodes. Our approach will remind some readers of Lavallée's (2007) methods for indirect sampling. The heart of the paper is Section 5, which sets forth unbiased estimators of population totals for cell-phone surveys and for corresponding dual-frame telephone survey designs. Section 6 gives an example, illustrating implications of the new methods of estimation for an existing telephone survey regarding the vaccination coverage of young children and teenagers. We close in Section 7 with a brief summary.

Throughout the article, we emphasize the development of rigorous but practical design and estimation procedures for population  $B$ . The methods of RDD surveys, *i.e.*, the methods for population  $A$ , are well known and, to a degree, have been used for decades; for a recent review of these methods see Wolter, Chowdhury and Kelly (2008).

## 2. Networks of units and the response protocol

In general, at least three types of units arise in the context of a cell-phone survey, as follows:

- Sampling units (SU)
- Reporting units (RU)
- Estimation units (EU).

The SU is the unit of sampling in the survey. In actual practice, telephone numbers may be sampled directly from cell-phone frames, or they may be sampled in stages, with perhaps exchanges or banks of numbers serving as the primary sampling units and numbers themselves being selected in one or more stages of subsampling within the primary units. To keep the discussion simple, in this article we will present the telephone number itself as the SU.



The actual target of the survey interview and the unit of analysis is what we shall call the EU. Some surveys focus on the collection and analysis of data on households or families, in which case the household or family is the EU. Other surveys focus on person level data, where the eligible persons may be children under age 18, adults age 18+, or some demographic segment of the population, such as Hispanic females aged 0-34. Still other surveys focus on both household- and person-level data, in which case the survey involves at least two types of EUs and two levels of analysis.

The adult is the respondent or RU in telephone surveys. The EU may or may not have the capacity to respond directly for itself, and instead an RU responds on its behalf. If the EU is an adult, then the same adult or even a different adult may serve as the corresponding RU. If the EU is a household, family, consumer unit, or child, then one or more adults may serve as the corresponding RU. The response protocol, specified by the survey methodologist, actually determines which RUs are permitted to respond for which EUs. In a typical survey, one respondent adult (or RU) would be contacted by telephone and interviewed for each SU selected into the sample.

SUs, RUs, and EUs may bear different relationships to one another in a cell-phone survey. Figure 1 gives nine networks that illustrate some of the types of relationships that are possible. In the first network, one SU is linked to one RU, which in turn responds for one EU. This arrangement could occur if one adult uses one telephone line, and the adult in turn reports for the household or for him or herself or for one child. In the second network, one SU is linked to two RUs, each of which can respond for the EU. This arrangement would occur, for example, if two adults shared the same telephone line and each was permitted by survey protocol to respond for the household. The fifth network could occur if two adults each had their own telephone line not shared with the other adult, while each adult in the pair is allowed by survey protocol to respond for each of two children.

More complicated networks are possible and surely must exist in the world. For example, the eighth network shows an arrangement of three adults sharing two telephone lines. The first of the lines is shared by all three adults, while the second line is only used by the third adult. The first of the adults is permitted by survey protocol to respond for two EUs, such as the adult's biological children; the second adult is not permitted to respond for any EUs; and the third adult is permitted to respond only for a third EU that is not reportable by the first two adults.

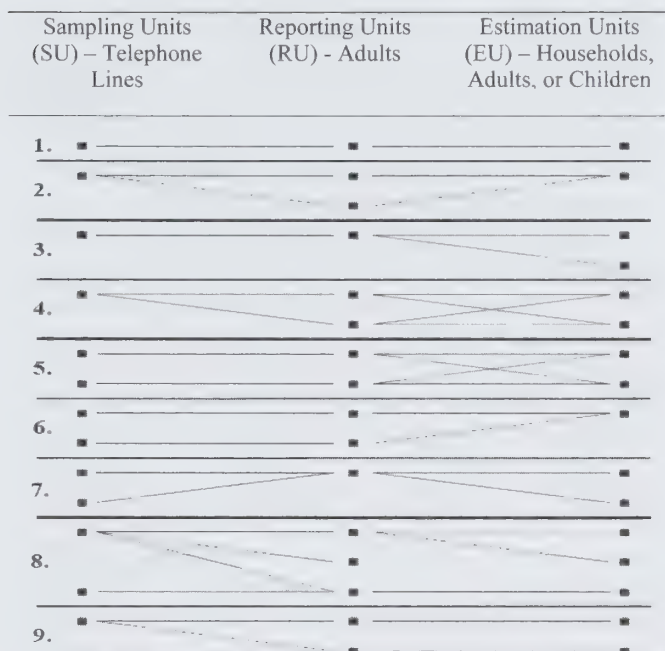


Figure 1 Examples of networks in a cell-phone survey

### 3. Links between units in the network

A *link* is a salient relationship between two nodes in the network. In the context of Figure 1, the links are represented by the line segments that join the different nodes. To provide a foundation for survey estimation, we need to explore links between (i) RUs and SUs, (ii) EUs and RUs, (iii) and EUs and SUs.

#### 3.1 Link of RU and SU

Two concepts are central to creating a link between an RU and an SU, namely, the concepts of (a) an *Active Personal Cell Number (APCN)* and (b) *usual access* to the cell-phone line.

An APCN is a telephone line that is in service at the time of the cell-phone survey and can ring through to an eligible adult who uses the cell phone, at least partially, for personal matters. In other words, an APCN meets three tests:

- It is in service
- It connects to an eligible adult respondent
- It is not used exclusively for business purposes.

We say that a given adult has usual access to a given APCN if and only if the individual has

- Regular,
- Substantial, and
- Ongoing use of the cell-phone line.

Each APCN has one or more regular adult users, and each individual user has usual access to one or more cell phones. In many cases, there is a unique one-to-one relationship between the cell-phone line and the adult user. In some cases, there is a one-to-many relationship between the cell-phone line and its users.

We treat a given SU and a given RU as linked if and only if the SU is an APCN and the RU has usual access to the SU. A cell-phone survey must work with and recognize the links that exist between the population of SUs and the population of RUs.

### 3.2 Link of EU and RU

A given EU is linked to one or more RUs via natural relationships that exist in the world, such as those created by family or place of residence. For example, an adult respondent may respond to the survey interview on behalf of his or her household, family, or consumer unit. He or she may respond for him or herself, for a dependent child under age 18, or for his or her own parent or sibling.

All surveys require a response protocol that defines which adult respondents are to respond for which EUs. The protocol is selected by the survey methodologist in light of feasibility, cost, and accuracy-of-reporting concerns. It is this protocol that establishes the links between EUs and RUs.

### 3.3 Link of EU and SU

The foregoing links between RUs and SUs and between EUs and RUs determine the links between EUs and SUs. We say a given EU is linked to a given SU if and only if the EU is linked to at least one RU that in turn is linked to the SU.

Some notation will become useful in our work in the following sections. Let  $j$  denote a given EU in the population of interest and let  $i$  be a given SU in the population. Then define the indicator or link variables

$$\ell_{ij} = \begin{cases} 1, & \text{if the } j^{\text{th}} \text{ EU is linked to the } i^{\text{th}} \text{ SU} \\ 0, & \text{otherwise.} \end{cases}$$

## 4. Duality between the populations of SUs and EUs

To begin the process of determining an unbiased estimation procedure for cell-phone surveys, we establish that a duality exists between the population of SUs or cell phones (henceforth denoted by  $U^{\text{SB}}$ ) and the population of EUs that are linked to cell phones (denoted by  $U^{\text{EB}}$ ). The goal of a cell-phone survey is to make inferences concerning  $U^{\text{EB}}$ , but we will soon see that this goal is equivalent to making certain inferences concerning  $U^{\text{SB}}$  (in this notation, the first

superscript designates the type of unit while the superscript  $B$  refers to the cell-phone sampling frame. Later we will use the superscript  $A$  to signify the landline sampling frame).

In the EU domain, a population total of interest is given by

$$Y^{\text{EB}} = \sum_{j \in U^{\text{EB}}} Y_j,$$

where the  $Y$ -variable on the right-hand side is a questionnaire item or other recoded or derived variable attached to the units in the population  $U^{\text{EB}}$ . Similarly, in the SU domain, a population total is defined by

$$X^{\text{SB}} = \sum_{i \in U^{\text{SB}}} X_i,$$

where the  $X$ -variable on the right-hand side is any fixed characteristic attached to the units in the population  $U^{\text{SB}}$ .

While the interest of the survey analyst centers on the total from the population of EUs (and on other parameters of this population), one can obtain a corresponding parameter in the SU domain by writing

$$Y^{\text{EB}} = \sum_{j \in U^{\text{EB}}} Y_j = \sum_{j \in U^{\text{EB}}} \sum_{i \in U^{\text{SB}}} \frac{Y_j \ell_{ij}}{\sum_{i' \in U^{\text{SB}}} \ell_{i'j}} = \sum_{i \in U^{\text{SB}}} X_i = X^{\text{SB}}, \quad (1)$$

where the  $X$ -variable is now defined specifically by

$$X_i = \sum_{j \in U^{\text{EB}}} \frac{Y_j \ell_{ij}}{\sum_{i' \in U^{\text{SB}}} \ell_{i'j}}. \quad (2)$$

From (1), one can see the correspondence between estimation in the SU domain and estimation in the EU domain. The total  $X^{\text{SB}}$ , with  $X_i$  defined as in (2), is equivalent to the total of interest  $Y^{\text{EB}}$ , and thus the problem of estimation of  $Y^{\text{EB}}$  is equivalent to the problem of estimation of  $X^{\text{SB}}$ .

We note that (2) arises in substantially the same form in the theory of indirect sampling. See Lavallée (2007), Theorem 4.1. In indirect sampling, SUs are linked to naturally defined clusters of EUs; if a given SU is selected into the sample, the survey data are collected for all EUs in the linked clusters. The analogy here is that the clusters are defined by the RUs that respond to the cell-phone interview attempt, and survey data are collected from the respondent for all EUs to which he or she is linked. The current situation is such that the cluster is defined by the SU-RU pair. An identifiability problem arises in this regard that does not occur in general in indirect sampling, and we elaborate on this matter in Section 5.5.

In (2), we effectively allocate an equal share of  $Y_j$  to each SU  $i$  to which it is linked. We could, alternatively, achieve the same ends by allocating  $Y_j$  to its linked SUs in proportion to some other known measure of the intensity of

the relationship between  $j$  and  $i$ . Although one could conceive of an optimal allocation of  $Y_j$  to its linked SUs, as in Deville and Lavallée (2006), such an allocation may be difficult to execute or may not be of great import in large scale practical settings.

## 5. Estimation

As mentioned in the introduction, some EUs will be linked exclusively to cell phones, some will be linked exclusively to landlines, and some will be linked to both landlines and cell phones. Phoneless EUs, if any, will not be linked to cell phones or to landlines. To provide notation for this environment, let  $U^E$  be the overall population of EUs of interest, and let  $U^S$  be the overall population of SUs. Let  $U^{EA}$  be the elements of  $U^E$  that are linked to landlines, let  $U^{EB}$  be the elements that are linked to cell-phone lines, let  $U^{Ea}$  be the elements that are linked only to landlines, let  $U^{Eb}$  be the elements that are linked only to cell-phone lines, let  $U^{Eab}$  be the elements that are linked to both landlines and cell-phone lines, and let  $U^{EC}$  be the elements that are phoneless. Note that  $U^E = U^{EA} \cup U^{EB} \cup U^{EC}$ ,  $U^{EA} = U^{Ea} \cup U^{Eab}$ , and  $U^{EB} = U^{Eb} \cup U^{Eab}$ , where  $U^{Ea}$ ,  $U^{Eb}$ , and  $U^{Eab}$  are disjoint sets. Also, let  $U^{SA}$  be the population of landlines, such that  $U^S = U^{SA} \cup U^{SB}$ . Landlines and cell-phone lines reflect disjoint subsets of the overall population of SUs.

In the following Sections 5.1 and 5.2, we discuss unbiased estimation for the subpopulation, say  $U^{ET} = U^{EA} \cup U^{EB}$ , that is linked to at least one telephone of any kind. We use the super-script  $T$  to designate this telephone subpopulation. Subsequently, in Section 5.4, we briefly discuss coverage of the phoneless population.

For EUs in  $U^E$ , define the indicator variables

$$\begin{aligned} \delta_j &= 1, && \text{if none of the RUs linked to } j \text{ have access} \\ &&& \text{to landline service, while at least one of} \\ &&& \text{these RUs has usual access to cell-phone} \\ &&& \text{service} \\ &= 0, && \text{otherwise} \\ \phi_j &= 1, && \text{if none of the RUs linked to } j \text{ have usual} \\ &&& \text{access to cell-telephone service, while at} \\ &&& \text{least one of these RUs has access} \\ &&& \text{to landline service} \\ &= 0, && \text{otherwise.} \end{aligned}$$

The  $\delta$ -variable is an indicator of cell-phone-only status and the  $\phi$ -variable is an indicator of landline-only status.

Then the population total of interest may be decomposed as

$$Y^{ET} = Y^{EA} + Y^{Eb}, \quad (3)$$

where

$$Y^{Eb} = \sum_{j \in U^{ET}} \delta_j Y_j$$

is the total of the cell-phone-only domain, and

$$Y^{EA} = \sum_{j \in U^{ET}} (1 - \delta_j) Y_j$$

is the total of the complement of this domain, including EUs that are linked exclusively to landlines and mixed EUs that are linked to both landlines and cell phones. The total of EUs may also be written as

$$Y^{ET} = Y^{Ea} + Y^{Eab} + Y^{Eb}, \quad (4)$$

where

$$Y^{Ea} = \sum_{j \in U^{ET}} \phi_j Y_j$$

is the total of the landline-only population, and

$$Y^{Eab} = \sum_{j \in U^{ET}} (1 - \delta_j) (1 - \phi_j) Y_j$$

is the total of the mixed population that has a combination of landline and cell-phone access. Finally, the population total may be written as

$$Y^{ET} = Y^{Ea} + Y^{EB}, \quad (5)$$

where

$$Y^{EB} = \sum_{j \in U^{ET}} (1 - \phi_j) Y_j$$

is the total of the complement (in the telephone population) of the landline-only population.

We view (3) and, to some extent, (4) as the decompositions of current practical interest and importance in telephone surveys in the USA and, in what follows, we present methods of estimation for each. Because of the current high relative cost of cell-phone interviews, surveys based on decomposition (5) would not be cost effective. It would almost always be better to represent the domain  $U^{Eab}$  using a sample of landlines than using a sample of cell phones. If the relative cost of cell-phone interviewing shifts downward in the future, decomposition (5) could become economically viable. It may also be viable for surveys in other countries where the cost structure is more favorable to cell-phone interviews.

### 5.1 Case of nonoverlapping domains

In this section, we will use a sample of cell-phone lines for purposes of estimation for the cell-phone-only population  $U^{Eb}$  and a sample of landlines for estimation for the entire landline population  $U^{EA}$ . We observe that it is not



possible to directly select a sample of cell-phone-only lines, because cell-phone-only status is not available on the sampling frame but rather is determined in the survey screening interview. To operationalize this design, one would screen-out cell-phone respondents who classify themselves in the mixed domain and terminate the interview, continuing the interview only for cell-phone-only respondents.

Let  $s^{SB}$  denote a probability sample of SUs (cell-phone lines) selected from the population  $U^{SB}$ , and let  $\{W_i^{SB}\}$  denote the set of base sampling weights such that

$$\hat{X}^{SB} = \sum_{i \in s^{SB}} W_i^{SB} X_i$$

is an unbiased estimator of the population total  $X^{SB}$ , where  $X_i$  is a characteristic of the  $i^{\text{th}}$  unit in the population. Assuming simple random sampling without replacement within strata, the base weights are of the form

$$W_i^{SB} = N_h / n_h, \quad (6)$$

where  $h$  signifies the sampling stratum in which the  $i^{\text{th}}$  SU is selected,  $N_h$  is the number of SUs on the sampling frame in stratum  $h$ , and  $n_h$  is the sample size in stratum  $h$ . Typically, the cell-phone sampling frame would include all telephone numbers within the exchanges assigned by the telephone system to cell phones. Simple random sampling would be the most common method of sample selection from such exchanges. There is little information available on the cell-phone sampling frame to enable stratification of the sample, except for the coarse geographic information embodied within the area code.

Let  $s^{EB}$  be the corresponding sample of EUs, *i.e.*,  $s^{EB} = \{j \in U^{EB} \mid j \text{ is linked to at least one SU } i \text{ in } s^{SB}\}$ . We will use this sample to estimate the domain total of EUs that are linked only to a cell phone,  $Y^{Eb}$ . From (1) and (2), we can readily see that the unbiased estimator of the domain total is given by

$$\begin{aligned} \hat{Y}^{Eb} &= \sum_{i \in s^{SB}} W_i^{SB} \left\{ \sum_{j \in U^{EB}} \delta_j Y_j \ell_{ij} / \sum_{i' \in U^{SB}} \ell_{i'j} \right\} \\ &= \sum_{j \in s^{EB}} \delta_j Y_j W_j^{EB}, \end{aligned} \quad (7)$$

where the EU level sampling weights are defined by

$$W_j^{EB} = \sum_{i \in s^{SB}} W_i^{SB} \ell_{ij} / \sum_{i' \in U^{SB}} \ell_{i'j} \text{ for } j \in s^{EB}. \quad (8)$$

Again, see Lavallée (2007) for expression of these weights in the context of indirect sampling.

Before leaving domain  $b$ , we observe in passing that it is possible to subsample the EUs and collect the survey information only for the subsample instead of enumerating all EUs linked to the sample RUs. If the statistician would

choose some form of subsampling, perhaps to control sample size or cost, then an additional weighting factor would appear in the weights in (8). Such subsampling is referred to as two-stage indirect sampling in Lavallée (2007, Section 5.1).

Turning to domain  $A$ , let  $s^{SA}$  denote a standard RDD sample of landline telephones, let  $s^{EA}$  be the implied sample of EUs, *i.e.*,  $s^{EA} = \{j \in U^{EA} \mid j \text{ is linked to at least one SU } i \text{ in } s^{SA}\}$ , and let

$$\hat{Y}^{EA} = \sum_{j \in s^{EA}} W_j^{EA} Y_j \quad (9)$$

be the standard unbiased estimator of the population total. For brevity, we shall not derive the standard sampling weights here; for more information about these weights, see Wolter *et al.* (2008).

From (7) and (9), the unbiased estimator of the population total of the EUs is given by

$$\hat{Y}^{ET} = \hat{Y}^{EA} + \hat{Y}^{Eb} \quad (10)$$

and the weights needed to support this estimator are  $\{W_j^{EA}\}$  and  $\{W_j^{EB}\}$ .

## 5.2 Case of overlapping domains

We now proceed with estimation starting from the decomposition (4). This means that in the cell-phone sample we will interview not only the cell-phone-only population, but also the mixed population (*i.e.*, those that use both landline and cell telephones). The estimator of the population total of interest is now of the form

$$\hat{Y}^{ET} = \hat{Y}^{Ea} + \hat{Y}^{Eab} + \hat{Y}^{Eb}, \quad (11)$$

where

$$\hat{Y}^{Ea} = \sum_{j \in s^{EA}} W_j^{EA} \phi_j Y_j$$

is the estimator for the landline-only domain derived from the landline sample,  $\hat{Y}^{Eb}$  is defined in (7) and is the estimator for the cell-phone-only domain derived from the cell-phone sample, and  $\hat{Y}^{Eab}$  is an estimator of the mixed domain obtained from both samples. The estimator of the mixed domain is

$$\begin{aligned} \hat{Y}^{Eab} &= \lambda \sum_{j \in s^{EA}} W_j^{EA} (1 - \phi_j) Y_j \\ &+ (1 - \lambda) \sum_{j \in s^{EB}} W_j^{EB} (1 - \delta_j) Y_j. \end{aligned} \quad (12)$$

The weights need to support estimator (11) are  $\{W_j^{EA}\}$  and  $\{W_j^{EB}\}$ .

See Hartley (1962) for discussion of the mixing parameter  $\lambda$  in a dual-frame survey, focusing on considerations

of sampling variability. Turning to considerations of bias, Brick *et al.* (2006) report that the propensity to respond to a cell-phone survey may be positively related to the frequency of use of the cell phone. Thus, the two pieces on the right side of (12) may be subject to a differential nonresponse bias not removed by the standard weighting-class methods. In the mixed population, infrequent users of the cell phone may be less likely to respond if surveyed in the cell-phone sample than if surveyed in the landline sample. If these adults would be substantially different from other adults in the mixed population with respect to the key characteristics under study in the survey, then (12) and also (11) could be subject to a nonresponse bias.

### 5.3 Variance estimation

To make inferences from the sample to the overall population, we require an estimator of the variance of the estimated total. First, consider the case of nonoverlapping domains. By working in the SU population, we can employ methods of variance estimation appropriate to the survey design. From (7), the estimated total for the cell-phone only domain may be written by

$$\hat{Y}^{Eb} = \sum_{i \in s_h^{SB}} W_i^{SB} X_i,$$

where

$$X_i = \sum_{j \in U^{EB}} \delta_j Y_j \ell_{ij} / \sum_{i' \in U^{SB}} \ell_{i'j}. \quad (13)$$

Assuming simple random sampling, the unbiased estimator of the variance of the estimated total is given by

$$v(\hat{Y}^{Eb}) = \sum_{h=1}^L N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} s_{xh}^2,$$

where

$$s_{xh}^2 = \frac{1}{n_h - 1} \sum_{i \in s_h^{SB}} \left( X_i - \frac{1}{n_h} \sum_{i' \in s_h^{SB}} X_{i'} \right)^2.$$

If we would ignore the finite population correction factor, which would be possible in almost any real telephone survey, the variance estimator becomes

$$v(\hat{Y}^{Eb}) = \sum_{h=1}^L \frac{n_h}{n_h - 1} \sum_{i \in s_h^{SB}} \left( W_i^{SB} X_i - \frac{1}{n_h} \sum_{i' \in s_h^{SB}} W_{i'}^{SB} X_{i'} \right)^2. \quad (14)$$

Now let  $v(\hat{Y}^{EA})$  be an estimator of the variance of  $\hat{Y}^{EA}$  for the RDD sample of landlines. Such estimators are well known and we do not review them here; see for example, Wolter *et al.* (2008). Because sampling is independent in the landline and cell-phone sampling frames, the unbiased

estimator of the variance of the estimated total for the entire telephone population becomes

$$v(\hat{Y}^{ET}) = v(\hat{Y}^{EA}) + v(\hat{Y}^{Eb}). \quad (15)$$

To facilitate the following developments, we let  $\hat{V}^{EB}[\delta Y]$  be another symbol to represent the estimator of variance in (14). This notation will emphasize the fact that the estimator of variance is based on the  $X_i$  variable in (13) defined in terms of the characteristic  $\delta_j Y_j$ , which is the characteristic of interest for cell-phone-only EUs. Also, let the symbol  $\hat{V}^{EA}[Y]$  be the estimator  $v(\hat{Y}^{EA})$  defined in terms of the characteristic  $Y_j$ . With this notation, (15) becomes  $v(\hat{Y}^{ET}) = \hat{V}^{EA}[Y] + \hat{V}^{EB}[\delta Y]$ .

Second, consider variance estimation for the case of overlapping domains. The estimator of the total of the telephone population is now  $\hat{Y}^{ET}$  in (11). For fixed  $\lambda$ , the unbiased estimator of variance is clearly seen from the work done in (14) and (15). It is

$$v(\hat{Y}^{ET}) = \hat{V}^{EA}[\phi Y + \lambda(1 - \phi)Y] + \hat{V}^{EB}[\delta Y + (1 - \lambda)(1 - \delta)Y]. \quad (16)$$

The first term on the right side of (16) is the variance estimator for the RDD sample of landlines applied to the composite characteristic  $\phi_j Y_j + \lambda(1 - \phi_j)Y_j$ , which is the characteristic for landline-only EUs plus a  $\lambda$ -portion of the characteristic for mixed EUs. The second term on the right side of (16) is the variance estimator for the cell-phone sample applied to the composite characteristic  $\delta_j Y_j + (1 - \lambda)(1 - \delta_j)Y_j$ , which is the characteristic for cell-phone-only EUs plus a  $(1 - \lambda)$ -portion of the characteristic for mixed EUs.

Estimators of covariance matrices can be built up from expressions like (15) and (16), facilitating statistical inference concerning other population parameters of interest.

### 5.4 Adjustments of the sampling weights

The sampling weights may be adjusted because of non-response or a planned calibration to known control totals.

Thus far, we have not addressed the various types of missing data that may occur in a cell-phone survey. We will focus on deriving adjustments for missing data that arise during the cell-phone interviews, assuming that standard adjustments for missingness in the landline sample have already been incorporated in the  $\{W_i^{EA}\}$  weights.

Missing data can arise due to three factors: (i) *non-resolution* of the SU; (ii) an *incomplete screening interview* of the RU; and (iii) an *incomplete main interview* of the RU. In this article, we adopt the convention that the resolution step refers to the classification of the SU as an ACPN or something else, such as a disconnected line or a dedicated business line; nonresolved SUs and SUs resolved as



non-ACPNs do not continue with the interview. The screening step refers to a brief preliminary interview intended to ascertain telephone status and to determine any demographic or other eligibility characteristics of any EUs linked to the RU; RUs for which the screening interview is incomplete or for which the screening interview is complete but no eligible EUs are linked to the RU do not continue with the interview. If the survey protocol calls for including only cell-phone-only EUs, as in Section 5.1, then the interview would terminate at this point for any mixed EUs. On the other hand, if the survey protocol calls for including both cell-phone-only and mixed EUs, as in Section 5.2, then the interview would continue for all such EUs. The interview step refers to the collection of the main survey items that form the substance of the survey for each of the eligible EUs linked to the RU. The survey methodologist must institute a definition of what constitutes a completed interview. In particular, the methodologist must decide whether *breakoffs* (an interview attempt that is completed for some but not all of the eligible EUs linked to the RU) are to be treated as a completed interview or not. Some other authors may organize the steps in the survey response process somewhat differently than the convention adopted here.

Adjustments to the sampling weights can be made for nonresolution and screener nonresponse, assuming a missing-at-random model for the response mechanism. These two adjustments must be made at the SU level. Let  $\{s_{\alpha}^{SB}\}$  be a partition of the cell-phone sample into user-specified weighting cells  $\alpha$ , and let the base sampling weights from (6) now be denoted by  $W_{1i}^{SB}$ , where the subscript 1 has been added simply to signify the first step in a multi-step adjustment process. Telephone area codes, rate centers, and census environmental variables at the county or area code level can be used to form the weighting cells; otherwise, little covariate information is available concerning cell-phone numbers. The cell-specific resolution completion rates are defined by

$$R_{1\alpha} = \frac{\sum_{i' \in s_{\alpha}^{SB}} r_{1i'} W_{1i'}^{SB}}{\sum_{i' \in s_{\alpha}^{SB}} W_{1i'}^{SB}},$$

where  $r_{1i}$  is a resolution indicator variable ( $= 1$ , if resolved,  $= 0$ , if not resolved), and the nonresolution adjusted weights are  $W_{2i}^{SB} = r_{1i} W_{1i}^{SB} / R_{1\alpha}$  for  $i \in s_{\alpha}^{SB}$ .

Let  $e_{1i}$  be an indicator of whether  $i$  is a resolved APCN ( $= 1$ , if resolved APCN,  $= 0$ , otherwise), and let  $\{s_{\beta}^{SB}\}_{\beta=1}^B$  be a partition of the cell-phone sample into user-specified weighting cells, which could be the same as or different than the foregoing partition. Then, the cell-specific screener completion rates are

$$R_{2\beta} = \frac{\sum_{i' \in s_{\beta}^{SB}} r_{2i'} e_{1i'} W_{2i'}^{SB}}{\sum_{i' \in s_{\beta}^{SB}} e_{1i'} W_{2i'}^{SB}},$$

where  $r_{2i}$  is a screener indicator variable ( $= 1$ , if screener completed,  $= 0$ , if screener not completed), and the screener-nonresponse adjusted weights are  $W_{3i}^{SB} = r_{2i} e_{1i} W_{2i}^{SB} / R_{2\beta}$  for  $i \in s_{\beta}^{SB}$ . Note that the appropriate sum of the weights is preserved at each step of the adjustment process.

Next, an adjustment to the sampling weights must be made for interview nonresponse. Depending on how break-offs are classified by the survey methodologist, there may be two cases to consider: (i) the RU completes or fails to complete the interview for all of its linked and eligible EUs en masse, or (ii) the RU selectively completes or fails to complete the interview on an EU by EU basis. If breakoffs would be classified as incomplete interviews, then only Case i would apply. Let  $e_{2i}$  be an indicator of whether the RU is screened and is linked to at least one EU that is eligible for the interview ( $= 1$ , if screened and eligible,  $= 0$ , otherwise), and let  $r_{3i}$  be the interview indicator variable ( $= 1$ , if the interview is complete,  $= 0$ , otherwise).

For Case i, the weight adjustment can be made at the SU level and is given by  $W_{4i}^{SB} = r_{3i} e_{2i} W_{3i}^{SB} / R_{3\gamma}$  for  $i \in s_{\gamma}^{SB}$ , where  $R_{3\gamma}$  is the weighted interview completion rate computed within user-specified weighting cells  $\gamma$ . Again, options for constructing weighting cells are limited in a cell-phone survey; they may be specified in terms of the information available at the previous weighting steps or any information collected in the screening interview. The weighted interview completion rate is

$$R_{3\gamma} = \frac{\sum_{i' \in s_{\gamma}^{SB}} r_{3i'} e_{2i'} W_{3i'}^{SB}}{\sum_{i' \in s_{\gamma}^{SB}} e_{2i'} W_{3i'}^{SB}}.$$

The estimated total for the cell-phone-only domain may now be expressed by

$$\hat{Y}^{Eb} = \sum_{j \in s^{1B}} \delta_j Y_j W_{4j}^{EB}, \quad (17)$$

where

$$W_{4j}^{EB} = \sum_{i \in s^{SB}} W_{4i}^{SB} \ell_{ij} / \sum_{i' \in U^{SB}} \ell_{i'j}$$

and  $s^{EB}$  is the set of eligible EUs reported in the screening interviews. The weight is zero for any eligible EUs in  $s^{EB}$  for which the RU failed to complete the main interview. The estimated total for the mixed domain, if called for by the survey protocol, is defined similarly by



$$\hat{Y}^{\text{Eab}} = \lambda \sum_{j \in s^{\text{EA}}} W_j^{\text{EA}} (1 - \phi_j) Y_j + (1 - \lambda) \sum_{j \in s^{\text{EB}}} W_j^{\text{EB}} (1 - \delta_j) Y_j.$$

For Case ii, the noninterview adjustment must be made at the EU level. The EUs are treated as spawned cases and a decision is made for each one as to whether it has a completed interview or not. The estimated total for the cell-phone-only domain is (17), where the weight is now defined by

$$W_{4j}^{\text{EB}} = r_{3j} e_{2j} W_{3j}^{\text{EB}} / R_{3\gamma} \quad \text{for } j \in s^{\text{EB}},$$

$$W_{3j}^{\text{EB}} = \sum_{i \in s_1^{\text{SB}}} W_{3i}^{\text{SB}} \ell_{ij} / \sum_{i' \in s_1^{\text{SB}}} \ell_{i'j},$$

and

$$R_{3\gamma} = \frac{\sum_{j' \in s_1^{\text{EB}}} r_{3j'} W_{3j'}^{\text{EB}}}{\sum_{j' \in s_1^{\text{EB}}} W_{3j'}^{\text{EB}}}.$$

Here, the weighting cells,  $\gamma$ , are defined in terms of characteristics of the EUs as determined from the screening interview and other sources.

For either Case i or ii, to facilitate computations, take  $W_{4j}^{\text{EA}}$  to be defined and equal to zero for EUs in the cell-phone sample, and take  $W_{4j}^{\text{EB}}$  to be equal to zero for EUs in the landline sample. If the survey protocol is as in Section 5.1, then we conclude that the survey weights for estimating the population total of interest are defined by

$$W_j = W_{4j}^{\text{EA}} + W_{4j}^{\text{EB}} \delta_j \quad (18)$$

for  $j \in s^{\text{ET}}$ , where  $s^{\text{ET}} \in s^{\text{EA}} \cup s^{\text{EB}}$ . Otherwise, if the survey protocol is as in Section 5.2, then we conclude that the survey weights are defined by

$$W_j = W_{4j}^{\text{EA}} \{\phi_j + \lambda(1 - \phi_j)\} + W_{4j}^{\text{EB}} \{\delta_j + (1 - \lambda)(1 - \delta_j)\} \quad (19)$$

for  $j \in s^{\text{ET}}$ .

The nonresponse-adjusted weights from (18) or (19) may be calibrated (Deville and Särndal 1992) to external control totals within socio-economic or geographic cells for the population of EUs, using poststratification, raking, or GREG (generalized regression estimation) techniques. If accurate sources are available, control totals may be established and calibration may be conducted separately for domains  $A$  and  $b$  or for domains  $a$ ,  $ab$ , and  $b$ . If control totals are not available by telephone status, then calibration must use control totals for the entire population regardless of telephone status.

To illustrate these ideas, we briefly examine the GREG estimator. Let us suppose that we have available a  $1 \times p$  auxiliary variable  $\mathbf{Z}_j$  for the observed, eligible EUs for which the control totals  $\mathbf{Z}^{\text{ET}} = \sum_{j \in U^{\text{ET}}} \mathbf{Z}_j$  are known. For example, the  $z$ -variable may arise from a fully saturated model in terms of explanatory variables age, race, and sex. Let  $s_4^{\text{ET}}$  be the set of EUs with a completed main interview and let  $n_4^{\text{ET}} = \#(s_4^{\text{ET}})$  be the number of eligible EUs reported in the completed interviews obtained within the consolidated telephone sample. Stack the  $y$ -values,  $z$ -values, and weights into the matrices  $\mathbf{Y} = (Y_1, \dots, Y_{n_4^{\text{ET}}})'$ ,  $\mathbf{Z} = (\mathbf{Z}'_1, \dots, \mathbf{Z}'_{n_4^{\text{ET}}})'$ , and  $\mathbf{W} = \text{diag}(W_1, \dots, W_{n_4^{\text{ET}}})'$ . Then the GREG estimator (Cassel, Särndal, and Wretman 1976) of the total of the telephone population of interest takes the familiar form

$$\hat{Y}^{\text{ET}} = \hat{Y}^{\text{ET}} + (\mathbf{Z}^{\text{ET}} - \hat{\mathbf{Z}}^{\text{ET}}) \hat{\beta} = \sum_{j \in s_4^{\text{ET}}} W_j g_j Y_j,$$

where the estimated coefficients are given by  $\hat{\beta} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{W}\mathbf{Y}$ ,  $\hat{Y}^{\text{ET}} = \sum_{j \in s_4^{\text{ET}}} W_j Y_j$ ,  $\hat{\mathbf{Z}}^{\text{ET}} = \sum_{j \in s_4^{\text{ET}}} W_j \mathbf{Z}_j$ , and  $g_j = 1 + (\mathbf{Z}^{\text{ET}} - \hat{\mathbf{Z}}^{\text{ET}}) \mathbf{Z}'_j$ . Lavallée (2007, Chapter 7) derives the Taylor series estimator of the variance of the GREG estimator in an indirect sampling context. Also see Wolter (2007, Chapter 6) for estimation of the variance of the GREG estimator.

Before leaving the topic of calibration, we note that we have largely left aside the small phoneless population, which fundamentally is impossible to sample in a telephone survey. Yet, in all likelihood, the overall population total  $Y^{\text{E}} = Y^{\text{ET}} + Y^{\text{EC}}$  will be the parameter of interest, not the total of the telephone population  $Y^{\text{ET}}$ , and the known control totals used in calibration may be totals for the overall population  $\mathbf{Z}^{\text{E}} = \mathbf{Z}^{\text{ET}} + \mathbf{Z}^{\text{EC}}$ , not totals for the telephone population  $\mathbf{Z}^{\text{ET}}$ . To include the phoneless population, we may consider use of a revised GREG estimator with  $g_j = 1 + (\mathbf{Z}^{\text{E}} - \hat{\mathbf{Z}}^{\text{ET}}) \mathbf{Z}'_j$ . This revision takes the same model for the phoneless population as for the telephone population. See Keeter (1995) and Chowdhury, Montgomery and Smith (2008) for other considerations in the calibration of weights for the phoneless population.

## 5.5 Identifiability assumptions

The foregoing theory assumes fundamentally that if SU  $i$  is selected into the sample of cell-phone lines, then  $X_i$  defined in (2) is observable in the cell-phone interview. Yet the 9<sup>th</sup> network (and also the 8<sup>th</sup>) in Figure 1 illustrates a potential problem for the theory. For this network, two RUs are linked to one SU, and in turn each RU is linked to only one EU. To continue this illustration, we suppose that these two EUs are not linked to any other RUs in the population. At the time of the survey interview, only one of the RUs will typically be reached and interviewed (unless the survey protocol would specifically mandate that an interview be

attempted with each RU linked to the selected SU). The respondent RU will report for its linked EU, but by the very nature of this network, the respondent cannot report for the EU that is linked to the companion RU who shares the sample cell-phone line. Thus, there is at least one EU that is linked to the SU that cannot be observed, *i.e.*, data cannot be collected in the cell-phone interview. Thus, we say  $X_i$  is *not identifiable*. The situation regarding the reportability of the two EUs would be reversed if the cell-phone interview attempt would have rung through to the companion RU.

To maintain the unbiasedness of the estimator of the population total, the  $X_i$  must be identifiable for every respondent SU selected into the sample of cell-phone lines. We need to make one of two assumptions. First, we could assume the problem away by acting as if networks like numbers 8 and 9 either do not exist or are trivial in number.

Secondly, the more realistic case would be to assume an extra randomization step, namely, that the interview call attempt to the given SU has reached a randomly selected RU linked to the SU. This randomization could be viewed as conceptual (that is, occurring naturally and not directed by the survey methodologist). To be formal and rigorous, one would need to collect information on the number of RUs linked to the SU and the probability that the cell-phone call attempt would ring through to the respondent RU. The probability would be approximated by the respondent's self-report of his or her share of use of the cell phone. If only one RU is linked to the SU, then this probability is 1.0 and clearly this simple value would not need to be collected in the interview once it is reported that there is only one RU. If two or more RUs are linked to the SU, then the probability or share to be collected is denoted by  $\tau_{ik}$  for RUs indexed by  $k$ , where  $\sum_{k \in U_i^{RB}} \tau_{ik} = 1$  and  $U_i^{RB}$  is the set of RUs that are linked to the  $i^{\text{th}}$  SU. With this additional information in hand, an unbiased estimator of

$$X_i = \sum_{j \in U^{EB}} \frac{\delta_j Y_j \ell_{ij}}{\sum_{i' \in U^{SB}} \ell_{i'j}}$$

is given by

$$\hat{X}_i = \sum_{j \in U^{EB}} \frac{1}{\sum_{i' \in U^{SB}} \ell_{i'j}} \sum_{k \in U_i^{RB}} \alpha_{ik} \frac{\delta_j Y_j \ell_{ij} \ell_{ikj}}{\tau_{ik} \sum_{k' \in U_i^{RB}} \ell_{ik'j}}, \quad (20)$$

where  $\alpha_{ik}$  is an indicator variable signifying whether the  $k^{\text{th}}$  RU was the realized respondent or not for the  $i^{\text{th}}$  SU in  $s^{SB}$  and

$$\begin{aligned} \ell_{ikj} &= 1, && \text{if SU } i \text{ is linked to RU } k \text{ which} \\ &&& \text{in turn is linked to EU } j \\ &= 0, && \text{otherwise.} \end{aligned}$$

The data are now identified and one can plug (20) into (7), giving the revised estimator

$$\hat{Y}^{EB} = \sum_{j \in s^{EB}} \delta_j Y_j W_{0j}^{EB} \quad (21)$$

with revised weights

$$W_{0j}^{EB} = \sum_{i \in s^{SB}} W_i^{SB} \frac{1}{\sum_{i' \in U^{SB}} \ell_{i'j}} \sum_{k \in U_i^{RB}} \alpha_{ik} \frac{\ell_{ij} \ell_{ikj}}{\tau_{ik} \sum_{k' \in U_i^{RB}} \ell_{ik'j}}. \quad (22)$$

As an approximation, one could take the RUs to be equal users of the cell phone, in which case  $\tau_{ik}$  would simply be the reciprocal of the number of RUs linked to the SU  $i$  for all RUs  $k$ . Adjustments for nonresponse and calibration to control totals would proceed as before.

Alternatively, the survey methodologist could call for a real randomization step, which would require that the interviewer make a roster of the RUs linked to the SU and select one at random, or a pseudo randomization step using the last birthday method. Such methods are probably not feasible at this time, due to the difficulty of gaining cooperation in cell-phone interviews.

## 5.6 Implications for data collection

Certain information must be collected in the survey interview in order to support the calculation of the estimators discussed here.

To support the use of  $\delta_j$ , the cell-phone survey must collect information to establish whether any of the RUs linked to the EU have access to a landline telephone. The respondent RU must report this information both for himself or herself and for other RUs that may be linked to the EU.

To support the use of  $\phi_j$ , the landline survey must collect information to establish whether any of the RUs linked to the EU have regular access to a cell phone. The respondent RU must report this information both for himself or herself and for other RUs that may be linked to the EU. This report may be quite straightforward in the event that the response protocol only links EUs to RUs within the same household. For more complicated response protocols, the report could be difficult to obtain.

To support the use of  $\sum_{i' \in U^{SB}} \ell_{i'j}$  in calculating the survey weights, the survey must collect information to establish how many SUs in the population are linked to the reported EU  $j$ . The respondent RU must be able to report the number of cell phones, including their own, that ring to an RU who is linked to the given EU.

If the estimator given in (21) and (22) would be used in order to identify all of the EUs, then additional information must be collected in the interview. The respondent RU must know and report the number of RUs, including themselves, that are linked to both the selected SU and the reported EU.



The respondent RU must also know and report their share of use of the cell phone on which the interview is completed or be able to say that use is approximately equal.

## 6. Example: The National Immunization Survey (NIS)

We illustrate the information that must be collected in the survey interview using the NIS, a survey of parents of children age 19-35 months and of teens age 13-17 years sponsored by the Centers for Disease Control and Prevention (CDC) for the purpose of monitoring vaccination coverage rates (*i.e.*, the proportion of children who are up-to-date with respect to the recommended vaccination schedule) in the USA. Data collection in the NIS occurs in two phases: an RDD telephone survey of households with landline telephones that have children or teens in the eligible age range, followed by a survey mailed to the vaccination providers of the age-eligible children. The sampling frame for the telephone survey phase of the NIS consists of all landline telephone numbers in 1+ banks in the USA. Cellular telephone numbers in dedicated cellular banks are currently not included in the NIS sampling frame. When a household with an age-eligible child is identified in the telephone survey, the interview is conducted with the adult in the household who is identified as the most knowledgeable about the vaccination status of the child (nearly always the mother or father). During the telephone interview, data are collected for each age-eligible child in the household, including the demographic characteristics of the child, demographic characteristics of the child's mother, and socio-economic characteristics of the child's household. At the end of the telephone interview, consent is asked to contact the child's vaccination providers. If consent is given, all vaccination providers named by the telephone interview respondent are contacted by mail to obtain the child's provider-reported vaccination history, which is used in statistical analysis to evaluate vaccination status. Smith, Hoaglin, Battaglia, Khare and Barker (2005) provide a detailed description of the statistical methods used by the NIS.

Because of the growth of the cell-phone-only population, the proportion of the NIS target population that is covered by the landline sampling frame has decreased in recent years. Using data from the National Health Interview Survey, Khare, Singleton, Wouhib and Jain (2008) estimate that about 18 percent of eligible children and 10 percent of eligible teens may be missing from the NIS sampling frame. To address the increase in cell-phone-only households in the NIS target population, cell-phone interviews could be added to the NIS.

For the NIS, the telephone number is the SU, the knowledgeable mother or father is the RU, and the age-eligible child is the EU. For the landline RDD or A sample, the parent is a resident of the household to which the sample landline number is assigned, while for the cell-phone or B sample, the parent has regular access to the cell phone to which the sample telephone number is assigned. Children are not subsampled in the NIS, but rather the knowledgeable parent reports for all of their age-eligible children who live in their home (but not for any children who may live elsewhere). These elements of the survey protocol establish the links between RUs and SUs and between EUs and RUs.

One comprehensive NIS design is to conduct estimation by way of nonoverlapping domains and decomposition (3). That is, the A sample is used to represent all children linked to a landline household and the B sample is used to represent all children linked to a cell-phone-only parent. We considered and rejected decompositions (4) and (5) due to considerations of cost and the potential for differential nonresponse bias in estimation for the mixed population.

To implement the estimator in (10), we determine whether the A-sample child is landline-only through use of the following three questionnaire items:

- A1. Next I have some questions about cell phones in your household. In total, how many working cell phones do you and your household members have available for personal use? Please don't count cell phones that are used exclusively for business purposes.
- A2. How many [of these] cell phones do [LIST ALL ELIGIBLE CHILDREN]'s parents and guardians usually use?
- A3. Of all the telephone calls that you and your family receive, are nearly all received on cell phones, nearly all received on regular phones, or some received on cell phones and some received on regular phones? (IF ASKED ABOUT INCLUDING BUSINESS CALLS: Please do not include any business-related calls in your answer).

For the cell-phone or B sample, we establish whether the child is cell-phone-only using the following two questions.

- B1. Do you have a landline in your household? (INTERVIEWER PROBE IF YES: Please do not include modem only lines, fax only lines, lines used just for a home security system, beepers, pagers, or the cell phone).
- B2. Thinking just about the landline home phone, not your cell phone, if that telephone rang and someone was home, under normal circumstances how likely is it that it would be answered? Would you say



extremely likely, somewhat likely, somewhat unlikely, or not at all likely?

We would use Question B2, due to Cantor, Brownlee, Zukin and Boyle (2008), to determine whether the landline is actually used for voice communications and thus whether the respondent is in the *ab* or *b* domain.

Also for the B sample, to determine the number of cell phones in the population that are linked to a given age-eligible child, we would use the following two questions:

- B3. Next, I have some questions about cell phones in your household. In total, how many working cell phones do you and your household members have available for personal use? Please do not count cell phones that are used exclusively for business purposes, and please include the number we called.
- B4. How many of these cell phones do [LIST CHILDREN]'s parents and guardians usually use? Please include the number we called.

Responses to questions A1-A3 and B1-B4 permit the calculation of survey weights and implementation of the unbiased estimator of the population total given in (10).

## 7. Summary

In this article, we used some theory of indirect sampling and network sampling to demonstrate a statistical framework for the design and analysis of cell-phone surveys. We exhibited an unbiased estimator of the population total with respect to estimation units linked to sampling units. By implication, this theory gives a means of constructing estimators of other population parameters that can be expressed as functions of totals. We illustrated the issues using the NIS, a telephone survey about young children and teens.

Information from the survey interviews is needed to classify estimation units into the cell-phone-only domain, the landline-only domain, or the mixed domain. Reporting error could result in misclassifications and undermine the unbiasedness of the estimator, as could survey nonresponse in the cell-phone and landline interviews.

## Acknowledgements

The authors thank associate editor for helpful comments.

## References

- Arthur, A. (2007). The birth of a cellular nation. *The Source*. Mediamark Research Inc. Available from: [http://www.mediamark.com/mri/TheSource/sorc2007\\_09.htm](http://www.mediamark.com/mri/TheSource/sorc2007_09.htm), 3.
- Blumberg, S.J., and Luke, J.W. (2008). Wireless substitution: Early release of estimates from the National Health Interview Survey. National Center for Health Statistics. Available from: <http://www.cdc.gov/nchs/nhis.htm>.
- Brick, J.M., Dipko, S., Presser, S., Tucker, C. and Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.
- Brick, J.M., Edwards, W.S. and Lee, S. (2007). Sampling telephone numbers and adults, interview length, and weighting in the California Health Interview Survey cell phone pilot study. *Public Opinion Quarterly*, 71, 793-813.
- Cantor, J., Brownlee, S., Zukin, C. and Boyle, J. (2008). Do We Need to Worry About Wireless Substitution in Public Opinion Polls about Health Reform. Presentation at the AcademyHealth 25<sup>th</sup> Annual Research Meeting, Washington, DC.
- Carley-Baxter, L., Peytchev, A. and Lynberg, M. (2008). Comparison of cell phone and landline surveys: A design perspective. Paper presented at the annual meeting of the American Association for Public Opinion Research, New Orleans, LA.
- Cassel, C.-M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.
- Chowdhury, S., Montgomery, R. and Smith, P.J. (2008). Adjustment for noncoverage of nonlandline telephone households in and RDD Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.
- CTIA (2008). Wireless Quick Facts. Available from <http://www.ctia.org/advocacy/research/index.cfm/AID/10323>.
- Deville, J.-C., and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, 32, 165-176.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Ehlen, J., and Ehlen, P. (2007). Cellular-only substitution in the United States as lifestyle adoption: Implications for telephone survey coverage. *Public Opinion Quarterly*, 71, 717-733.
- Frankel, M., Battaglia, M., Link, M. and Mokdad, A. (2007). Integrating cell phone numbers into Random Digit-Dialed (RDD) landline surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, (Alexandria, VA), 3793-3800.
- Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.
- Keeter, S. (1995). Estimating non-coverage bias from a phone survey. *Public Opinion Quarterly*, 59, 196-217.
- Khare, M., Singleton, J.A., Wouhib, A. and Jain, N. (2008). Assessment of Potential Bias in the National Immunization Survey (NIS) from the Increasing Prevalence of Households Without Landline Telephones. Presented at the National Immunization Conference, Centers for Disease Control and Prevention.
- Lavallée, P. (2007). *Indirect Sampling*. New York: Springer Science+Business Media, LLC.

- Lavrakas, P.J., Shuttles, C.D., Steeh, C. and Fienberg, H. (2007). The state of surveying cell phone numbers in the United States: 2007 and Beyond. *Public Opinion Quarterly*, 71, 840-854.
- Smith, P.J., Hoaglin, D.C., Battaglia, M.P., Khare, M. and Barker, L.E. (2005). Statistical methodology of the National Immunization Survey, 1994-2002. National Center for Health Statistics, Hyattsville, MD. *Vital and Health Statistics*, Series 2, 138.
- Wolter, K.M. (2007). *Introduction to Variance Estimation, Second Edition*. New York: Springer-Verlag.
- Wolter, K.M., Chowdhury, S. and Kelly, J. (2008). Design, conduct, and analysis of random digit dialing surveys. In *Handbook of Statistics: Sample Surveys, Theory, Methods and Inference*, (Eds., D. Pfeffermann and C.R. Rao), Elsevier, Oxford, UK.





## Collecting data for poverty and vulnerability assessment in remote areas in Sub-Saharan Africa

Rudolf Witt, Diemuth E. Pems and Hermann Waibel<sup>1</sup>

### Abstract

Data collection for poverty assessments in Africa is time consuming, expensive and can be subject to numerous constraints. In this paper we present a procedure to collect data from poor households involved in small-scale inland fisheries as well as agricultural activities. A sampling scheme has been developed that captures the heterogeneity in ecological conditions and the seasonality of livelihood options. Sampling includes a three point panel survey of 300 households. The respondents belong to four different ethnic groups randomly chosen from three strata, each representing a different ecological zone. In the first part of the paper some background information is given on the objectives of the research, the study site and survey design, which were guiding the data collection process. The second part of the paper discusses the typical constraints that are hampering empirical work in Sub-Saharan Africa, and shows how different challenges have been resolved. These lessons could guide researchers in designing appropriate socio-economic surveys in comparable settings.

Key Words: Socio-economic household surveys; Survey design; Data collection challenges; Sub-Saharan Africa.

### 1. Introduction

To collect economic data in small-scale fisheries in Sub-Saharan Africa (SSA) is challenging, as patterns and constraints of resource use vary considerably, *i.e.*, spatially, seasonally and over time. This requires careful planning of the collection of data that is needed for meaningful poverty and vulnerability assessment. Although small-scale fisheries (SSF) can generate significant profits and make considerable contributions to poverty alleviation and food security, little information exists about their actual contribution to livelihoods and household economics in Sub-Saharan Africa (FAO 2005, 2006). The key constraints for empirical studies in this field are difficulties associated with data collection, such as remoteness and inaccessibility especially during the rainy season. High variability of natural resource conditions, and thus production, cause additional requirements for survey design. For preparation and implementation of a survey in SSA, researchers can draw upon similar studies in other parts of the world concerning survey methodology, questionnaire design, and interview procedure, *e.g.*, the World Bank's Living Standard Measurement Survey (LSMS) questionnaire. However, many peculiarities of rural communities in SSA require an adapted and elaborated approach.

Some of these peculiarities are of an ecological nature, such as seasonal changes in access to resources and markets, which are directly affecting patterns and constraints of resource use. Others pertain to the economic side of household behavior, since income-generating activities of rural households in SSA compose complex portfolios.

Particularly households in fishery-dependent communities have adopted a flexible and strongly seasonal matrix of diversified activities (Béné, Neiland, Jolley, Ovie, Sule, Ladu, Mindjimba, Belal, Tiotsop, Baba, Dara, Zakara and Quensiere 2003a; Béné, Neiland, Jolley, Ladu, Ovie, Sule, Baba, Belal, Mindjimba, Tiotsop, Dara, Zakara and Quensiere 2003b; Béné, Mindjimba, Belal, Jolley and Neiland 2003c; Neiland, Jaffry and Kudasi 2000, Neiland, Madaka and Béné 2005; Sarch 1997). The local populations are alternatively or simultaneously fishers, herders, and farmers, and each piece of land is potentially a fishing ground, a grazing area and a cultivated field, depending on the flood cycle (Béné *et al.* 2003a, page 20). Due to high vulnerability of the ecological and economic system to shocks, such as flood, drought and pest outbreaks which result in year to year variation in fish stocks and in high crop losses, households have diversified their activities portfolio, thus spreading the risk of income losses. Capturing the dynamic interplay of the different livelihood elements is a special challenge in conducting socio-economic household surveys. Other constraints for data collection are culturally determined, for example tensions between different ethnic groups, the existence of a multitude of languages and patois spoken in the study region, or some peculiarities of the Muslim-African culture.

The data required for poverty and vulnerability assessment demand an appropriate survey methodology, for data quality to meet the requirements of a robust econometric analysis. Data needs for economic poverty assessment and the evaluation of SSF's contribution to poverty and vulnerability alleviation are substantial. Detailed information on

1. Rudolf Witt and Hermann Waibel, Institute of Development and Agricultural Economics, Faculty of Economics and Management, Leibniz University Hannover, Königsworther Platz 1, 30167 Hannover, Germany. E-mail: witt@ifgb.uni-hannover.de; Diemuth E. Pems, Economist, The WorldFish Centre, Penang, Malaysia.

household income, including different income sources such as agricultural production, fishing, livestock rearing, off-farm work *etc.*, is necessary. Also, data on the stock and value of productive and convertible assets, as well as on the distribution of consumption expenditures need to be elicited. In addition, information on control variables, such as ecological, economic or social shocks that have occurred in the past, subjective risk assessments, debts and liabilities, household composition, and others, is required.

This paper presents the collection procedure of quantitative household data from poor households in the Logone floodplain, a major inland fisheries region in Northern Cameroon. The objective of collecting household level panel data in 2007-2008 was to assess the role of small-scale fisheries (SSF) in mitigating risk through portfolio diversification, thus contributing to reducing vulnerability to poverty. In this paper, we emphasize the requirements of the general methodological approach for sampling and survey design. Due to the complex nature of the SSF sector outlined above, a procedure for sampling and data collection is required that allows the assessment of poverty and vulnerability of SSF households. Particularly, the survey design needs to account for the high variation in income generating activities over time as a result of the high variability of access to natural resources and resulting adjustments in a household's food security situation, consumption, income and assets.

## 2. Study site and sampling procedure

The study site is the Logone floodplain in the Far-North province of Cameroon. The floodplain covers about 8,000 km<sup>2</sup> and is part of the bigger Logone-Chari subsystem in the Lake Chad Basin, which supplies 95% of Lake Chad's total riverine inputs and has a basin area of approximately 650,000 km<sup>2</sup> (UNEP 2004). Within this vast area a representative region was defined in collaboration with national experts and other key informants, while considering the accessibility and logistic feasibility of the study. The study area covers about 2,400 km<sup>2</sup>, spreading from the Maga Lake in the south to Ivyé village in the north, where the Logomatya joins the Logone River. This area is relatively densely populated and is characterized by rich fish stocks and intensive fishing, fish processing and fish trading.

The livelihoods of the rural population in this area are particularly exposed to harsh climatic conditions, such as limited and erratic rainfall, which result in a large variation of production outcomes from year to year (In this respect, the study area is representative for many similar rural settings, particularly in the Sudano-Sahelian zone of Sub-Saharan Africa.) and thus considerable income risk. However, the impact is different between the sub-regions of

the study area. Based on Neyman (1938), as cited in Rao (2005), a stratified random sampling procedure was therefore considered most effective. To draw a representative sample of households in the study area while accounting for different production conditions (such as access to fish resources), a stratification of the study site into different agroecological zones was undertaken. It was assumed that under different ecological and production conditions the role of fisheries in terms of income generation would differ. This procedure allowed capturing the whole continuum of fishing intensity (from specialized/full-time fishermen to purely agriculture/livestock rearing oriented households).

In a second step, a complete list of villages in the study area ( $N = 88$ ) was compiled. These villages served as the primary sampling unit. Following the recommendations of local fisheries experts, 14 villages were selected proportional to the total number of villages per zone. The average village size in the floodplain (study area) is about 45 households, with a range of 15 to 100 households. Within villages every second household was chosen randomly from household lists established by the village headman. Hence, a sample size of 300 households was chosen proportional to the size of the village populations, which equates to a sampling ratio of 7% of the total population (estimated at 20,000 by the Ministry of Livestock, Fisheries and Animal Industries, MINEPIA).

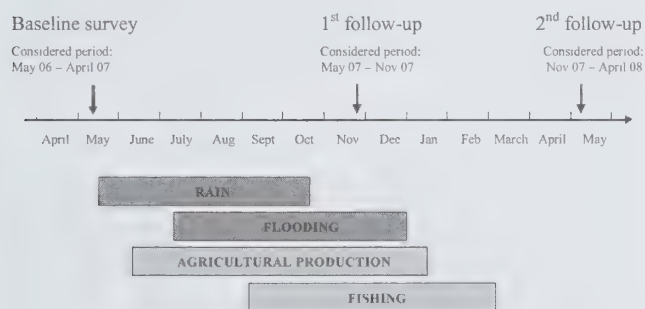
All selected villages were visited before commencing the household level survey with the aim to establish contacts between the researcher and the village headmen and conduct focus group discussions (FGDs) with the village leaders. The objective of the FGDs was twofold. First, some general information was collected such as the village size, infrastructure, and access to fish resources and markets. Second, complete household lists for every selected village were compiled, since no official statistical information existed. For this study, a household was defined as an economically independent unit consisting of the household head, one or more spouse(s), children and other directly dependent members, living in the household or having migrated to other locations. Household size varies from two (*i.e.*, normally husband and spouse) to more than 15. Large households are common for Northern Cameroon, since due to widespread polygamy household heads often live together with up to four wives. Mostly, households do not live separately from other kin households, but usually form a clan, living together in a larger compound. However, within the compound, households are independent from each other. During the visits, special attention was paid to list the names of individual household heads and not only those of the compound/clan leaders. The additional information collected during the FGDs was necessary to get a first understanding of the livelihood options and constraints in



the study area, which proved to be helpful for the development of the household questionnaire. In the last step, the compiled household lists were used for a weighted random sampling of the 300 sample households.

### 3. Survey design

Seasonality is an important characteristic of the livelihood conditions in the Logone floodplain. Therefore, in order to capture seasonal variation, the survey was designed to yield a two-period panel data set (2006 – 2007), with an additional third survey six months after conducting the baseline survey (see Figure 1). The baseline survey was accomplished right at the end of the dry season, when income-generating activities are extremely limited, and the financial resources, generated during the rainy season in 2006, are being used up. The period covered in the baseline survey was May 2006 to April 2007, constituting a stock check of average income flows, consumption expenditures, and an asset inventory. The first follow-up survey captured the busy time of the year, where expenditures rise due to investments (e.g., purchase of new fishing nets and other productive assets), and variable production costs in agriculture and fishing. Finally, the second follow-up survey covered the second half of the year, giving account of the economic household activities in this period. This approach was chosen to improve the accuracy of data on livelihood activities by reducing the recall period, and to make sure to capture seasonal variation in income and consumption.



Source: own illustration

**Figure 1** Livelihood options in the study area and design of the survey

Before the start of each survey, enumerator training workshops of 3 to 4 days were conducted, including pre-testing of the questionnaire in order to detect weaknesses and the necessity to eliminate, rephrase or add additional questions. The baseline pre-test was carried out in two villages of zone 1 and 2, in order to test the suitability of the questionnaire for different livelihood conditions. The baseline study was completed within 3 weeks in May 2007 by four enumerators, working in a team, and accompanied

and directly supervised by the first author. This procedure gave the opportunity for immediate cross-checking for missing information, and also enabled the researcher to observe and reinforce interview techniques and immediately discuss problems or questions.

Due to the relative remoteness of the villages and difficulties of access, careful logistical planning was necessary. The field trips often covered several days, and it was inevitable to spend the nights in the villages. Hence, the survey procedure adopted was as follows: the whole team arrived in a village, presenting itself to the village chief, who had been previously informed about the arrival date of the team during the FGD visit. The chief then called the heads of the selected households to a central meeting place, usually under a tree in front of the chief's house. After the interview, which normally took about one hour, the respondent was given a small present as a compensation for his time (a package of sugar and a bag of tea), and the next household head was called to sit down. Working in a group enabled the team to finish a village in about one or two days and proceed to the next one. That course of action strongly motivated and encouraged the enumerators for security and psychological reasons. The interview time, and hence the time planned to be spent per village, was held flexible, so that careful cross-checking for consistency and plausibility of responses was ensured. Hence, during the enumerator training workshops and throughout the data collection process, special emphasis was placed on the ultimate primacy of data quality.

### 4. Data collection challenges and lessons learnt

This section describes some challenges and constraints in data collection, which have been encountered during this study, but which are not limited to the study region. Similar settings are found in many wetlands and floodplains in SSA, and the lessons learnt in this study may prove helpful for comparable data collection endeavors.

#### *Seasonality*

When collecting data in rural fisheries-dependent communities in SSA, the seasonal nature of the livelihood systems and the ecological constraints need to be taken into consideration. Very often, villages are spatially marginalized and access is extremely difficult during certain periods of the year. For example, in the Logone floodplain in North Cameroon, access to the villages is very restricted during several weeks twice a year due to the annual flood cycle. At the beginning of the flooding season, and during the deflooding period, access is not possible, neither by vehicle, nor by boat. Hence, the placing of the survey periods need to be adapted to these conditions. For example,



although it would have been more reasonable to place a follow-up survey at the end of the production cycle in January, thus better capturing agricultural production and fishing harvests, this procedure proved to be unfeasible. From mid December to end of February access to the sampled villages was not possible at all. The research team decided for a compromise, collecting data in December, even if this falls in the midst of the harvesting season. The missed data on yields and income was then recollected during the second follow-up. Similar problems arise in other major inland fisheries such as the Hadejia-Nguru Wetlands in Nigeria or the Lower Shire river basin in Malawi.

### *Defining time periods*

For recall surveys and particularly for panel surveys (*i.e.*, the research team is repeatedly revisiting the same households) it is important to assure a common understanding of the time period that is considered in the questionnaire. Different notions of the time span may result in biased information concerning income or consumption flows and can flaw the results and conclusions drawn from the study. In order to assure a common understanding of the requested time period, the respective cultural understanding of time needs to be taken into account. We found that in the Logone floodplain, people do not think in time units such as weeks or months. Hence, questions, such as: "How much did you spend on food items in the last 6 months?" were not appropriate. In this case, it proved instrumental to refer to certain region-wide acknowledged social events or celebrations. For example, the survey in November coincided with the Tabaski festivities, so that it was easy for the respondents to delimit the time period considered in the second follow-up survey.

### *Selection of enumerators and their cultural competence*

Perhaps the most important factor in empirical work is the choice of the enumerators. To achieve good data quality, enumerators must not only provide the needed skills and knowledge, but also dispose over additional soft skills, such as mastering of languages, social competence, and the will to work under severe conditions.

The lack of sufficiently educated interviewer personnel in the Far-North Province in Cameroon presented a serious constraint. For this study, a team of five MINEPIA staff, who work as government officials in the survey area, was recruited as enumerators. While respondents can have reservations to provide information to government officers, the more important factor was that the survey team represented the two ethnic groups of the study area. Also, enumerators spoke the languages of the region, they were familiar with the local peculiarities, and used to the conditions in the field. In addition, respondents' willingness

to provide information was actually encouraged in expectations of a follow-up governmental support.

Another advantage of the selected enumerators was awareness and sensitivity towards ethnic tensions. Enumerators were careful not to take sides with either one of the involved parties, and avoided offensive statements. This was especially important with regard to multiple visits of villages and respondents during the follow-up surveys. Any disaccord between respondents and enumerators would have resulted in significant attrition and the need to drop entire villages from the sample.

Certain cultural or religious norms also demanded tactfulness and respect. For example, in a number of villages only men could be interviewed since women in that African-Muslim culture are not allowed to meet or talk to men other than direct family members. In cases where the household head was not present at the time of the visit, it was not possible to interview the spouse (or any other woman in the household) instead. An adult male household member had to be chosen to provide the required information. For the same reason, interviews could not take place in the house of the respondents. For the sake of compliance to these cultural norms, the interview procedure had to be adapted. Instead of visiting the chosen households one by one, all sampled household representatives in each village were called to a central meeting place by the village chief (usually in front of the chief's house). If the household head was not present, another adult member of the household (usually male) was interviewed. The enumerators then seated themselves at a distance of about three to five meters from each other, calling the respective respondent to be interviewed in private, while the others were waiting for their turn.

### *Sample attrition*

A particular challenge of panel surveys in general is to maintain the size of the sample over time (Jäckle and Lynn 2008, Laaksonen 2007). Attrition can be high due to several reasons. For example, in some cases the household head has died, the whole household has moved away, or the respondents lose interest to participate especially if no or not enough incentives are provided. The loss of willingness to participate in a follow-up survey caused a problem during the second visit. Due to budget constraints the survey team decided not to compensate the participants for their time at the second visit. For the baseline survey, each respondent had received a box of sugar and a package of tea which turned out to be a strong extrinsic incentive. When households learned that no remuneration had been foreseen at the second visit, 69 households (23% of the total sample) announced that they were "too busy" to participate. Considering this reaction, compensation was again offered at the

third survey, so that most of the lost households could be regained. They were even willing to respond to both questionnaires (1<sup>st</sup> and 2<sup>nd</sup> follow-up). Thus the missing data could be completed during the last survey round albeit at the cost of lower reliability due to memory bias. Such respondent behavior is consistent with findings by Jäckle and Lynn (2008), who report significant positive effects of continued incentive payments on attrition, bias and item non-response. At the end of the survey period, 14 households (4.7%) have been lost due to permanent migration or other reasons, and hence were removed from the sample.

## 5. Summary and conclusions

Data collection for poverty analysis in SSA is a challenging endeavor. Often, cultural, ecological and economic constraints push researchers to put up with a compromise between data quality and feasibility of the study. On the other hand, collection of such data is important because little is known about poverty and vulnerability of marginalized groups such as fisheries communities in remote areas of SSA. In this paper, we present the approach that has been taken in the course of a study on poverty and vulnerability in the Logone floodplain, which is a major fishing area in Northern Cameroon. We identify typical constraints that are often hampering empirical work in SSA, and show how different challenges can be overcome by an adequate survey design, sampling and careful application of the survey instrument. Major constraints encountered were the difficulties to access the target population, limitations in finding qualified enumerators and high demand for cultural sensitivity of the research team.

Of eminent importance is a close collaboration with local authorities and experts in the respective field of research, as well as a good understanding of and compliance with local cultural norms and values. Learning from the local population and empathizing with its peculiar ways of living before starting the survey per se has been found to be a key success factor for working in that region. Summing up, it can be concluded that despite a number of difficulties, quantitative data collection in rural Sub-Saharan Africa is a task that can be completed with satisfying results. An appropriate survey design and interview procedure developed in collaboration with local staff and experts can assure adequate data quality for economic poverty and vulnerability analysis.

## Acknowledgements

We thank the German Federal Ministry of Economic Cooperation and Development (BMZ) for the financial

support through the project on Food security and poverty alleviation through improved valuation and governance of river fisheries in Africa which was coordinated by the WorldFish Center. The authors also thank two anonymous referees for their very thorough review of the paper and their extremely valuable comments. We also want to thank the editor for the excellent guidance of the review process and additional useful comments on the paper. The views expressed in this paper are not necessarily those of the donor agency nor of our project partner the WorldFish Center.

## References

- Béné, C., Neiland, A., Jolley, T., Ovie, S., Sule, O., Ladu, B., Mindjimba, K., Belal, E., Tiotsop, F., Baba, M., Dara, L., Zakara, A. and Quensiére, J. (2003a). Inland fisheries, poverty, and rural livelihoods in the Lake Chad Basin. *Journal of Asian and African Studies*, 38, 1, 17-51.
- Béné, C., Neiland, A., Jolley, T., Ladu, B., Ovie, S., Sule, O., Baba, M., Belal, E., Mindjimba, K., Tiotsop, F., Dara, L., Zakara, A. and Quensiére, J. (2003b). Natural-resource institutions and property rights in inland African fisheries - The case of the Lake Chad Basin region. *International Journal of Social Economics*, 30, 3, 275-301.
- Béné, C., Mindjimba, K., Belal, E., Jolley, T. and Neiland, A. (2003c). Inland fisheries, tenure systems and livelihood diversification in Africa: The case of the Yaéré floodplains in Lake Chad Basin. *African Studies*, 62, 2, 187-212.
- FAO (2005). Technical guidelines for responsible fisheries Nr 10: Increasing the contribution of small-scale fisheries to poverty alleviation and food security. FAO, Rome.
- FAO (2006). FAO's Activities on Small-scale Fisheries: An Overview. Advisory Committee on Fisheries Research (ACFR), Sixth Session, Rome, 17-30 October 2006.
- Jäckle, A., and Lynn, P. (2008). Respondent incentives in a multi-mode panel survey: Cumulative effects on nonresponse and bias. *Survey Methodology*, 34, 1, 105-117.
- Laaksonen, S. (2007). Weighting for twophase surveyed data. *Survey Methodology*, 33, 2, 121-130.
- Neiland, A.E., Madaka, S.P. and Béné, C. (2005). Traditional Management Systems, poverty and change in the Arid Zone Fisheries of Northern Nigeria. *Journal of Agrarian Change*, 5, 117-48.
- Neiland, A.E., Jaffry, S. and Kudasi, D.K. (2000). *Fishing Income, Poverty and Fisheries Management in North-East Nigeria*. Fisheries of North East Nigeria and the Lake Chad Basin, Volume I - A compilation of research project reports in two volumes 1993-2004, compiled by Dr. Arthur E. Neiland, 291-319.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- Rao, J.N.K. (2005). Interplay Between Sample Survey Theory and Practice: An Appraisal. *Survey Methodology*, 31, 2, 117-138.

- Sarch, M.-T. (1997). Fishing and Farming in Lake Chad: Implications for Fisheries Development. *Development Policy Review*, 15, 141-57.
- UNEP (2004). Map of the Lake Chad Basin. In *Lake Chad Basin, GIWA Regional assessment 43*, (Eds., M.P. Fortnam and J.A. Oguntola), University of Kalmar, Kalmar, Sweden.



## Respondent differences and length of data collection in the Behavioral Risk Factor Surveillance System

Mohamed G. Qayad, Pranesh Chowdhury, Shaohua Hu and Lina Balluz<sup>1</sup>

### Abstract

The current economic downturn in the US could challenge costly strategies in survey operations. In the Behavioral Risk Factor Surveillance System (BRFSS), ending the monthly data collection at 31 days could be a less costly alternative. However, this could potentially exclude a portion of interviews completed after 31 days (late responders) whose respondent characteristics could be different in many respects from those who completed the survey within 31 days (early responders). We examined whether there are differences between the early and late responders in demographics, health-care coverage, general health status, health risk behaviors, and chronic disease conditions or illnesses. We used 2007 BRFSS data, where a representative sample of the noninstitutionalized adult U.S. population was selected using a random digit dialing method. Late responders were significantly more likely to be male; to report race/ethnicity as Hispanic; to have annual income higher than \$50,000; to be younger than 45 years of age; to have less than high school education; to have health-care coverage; to be significantly more likely to report good health; and to be significantly less likely to report hypertension, diabetes, or being obese. The observed differences between early and late responders on survey estimates may hardly influence national and state-level estimates. As the proportion of late responders may increase in the future, its impact on surveillance estimates should be examined before excluding from the analysis. Analysis on late responders only should combine several years of data to produce reliable estimates.

Key Words: BRFSS; Responders; Differences; Length of data collection.

### 1. Introduction

The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based household telephone survey in the United States (U.S.) and its territories which monitors health risk behaviors and chronic disease conditions for the adult noninstitutionalized population (Centers for Disease Control and Prevention [CDC] 2009a, BRFSS Turning Information into Public Health, <http://www.cdc.gov/brfss/about.htm>). It is the largest telephone survey in the world and is implemented by the 50 states, the District of Columbia, and U.S. territories, in collaboration with the CDC. The survey is conducted continuously throughout the year.

CDC dispenses the samples (phone numbers) to states quarterly. At the state level, the samples are divided into 12 monthly lists for operational purposes. Trained interviewers call each sampled telephone number. After each call to a sampled telephone number, a disposition code is assigned. States and their contractors are required to give final dispositions to their monthly released samples within that month. Over 90% of the monthly samples and completed interviews receive final dispositions within 31 days. States continue to complete their remaining samples afterwards (Qayad, Balluz and Garvin 2009).

Because of economic downturns, states and survey organizations may face budget cuts that could adversely affect their survey operations. Such unforeseen circumstances warrant searching for alternative operational strategies. A

cost-effective alternative could be to end data collection at the end of each month. However, ending data collection within one month excludes interviews completed after 31 days. Such exclusion could influence the variability of the respondents, surveillance estimates and the size of completed interviews, which could affect other operational decisions. Currently, the size of late responders is small and may not influence surveillance estimates. However, the current trend in survey responses heralds a continuous decline in survey responders, which could prolong the duration to reach respondent and the eventual increase in the proportion of late responders. Such circumstances require thorough examination of the influence of late responders on surveillance estimates in the future. This study examines whether respondents who completed the interviews within 31 days and those who completed after 31 days are different in demographics, risk behaviours, and chronic disease conditions.

### 2. Methods

We used the 2007 BRFSS data, which is an ongoing state-based random digit dialing (RDD) telephone survey among the non-institutionalized civilian population in the US. We divided the duration of the interview into two periods, 0-31 days and >31 days. Respondents who completed the interviews within 31 days (referred as early responders) and those completed after 31 days (referred as late responders).

1. Mohamed G. Qayad, Pranesh Chowdhury, Shaohua Hu and Lina Balluz, Division of Adult and Community Health, Behavioral Surveillance Branch, Centers for Disease Control and Prevention, Atlanta, GA 30341, U.S.A. E-mail: [maq3@cdc.gov](mailto:maq3@cdc.gov).

Demographic factors included were - gender, race, income and age. Race had four groups - white non-Hispanic, Black non-Hispanic, Hispanic and other race. Education had three levels: not a high school graduate, high school graduate, and more than high school education. Income categories were <\$15,000, \$15,000 - \$34,999, \$35,000 - \$49,999 and \$50,000 or more. Age had the following categories: 18 - 24 years, 25 - 44 years, 45 - 64 years, and 65 or more years. Respondents <65 years old who did not have any health plan (including health insurance, prepaid plans such as HMOs, or government plans such as Medicare) were considered not to have health plan. General health was dichotomized into good health (excellent, very good, or good health) and fair or poor health.

Health risk behaviors included were - binge drinking, current smoking, (lack of) physical activity, and (insufficient) fruit and vegetable consumption. Binge drinking was defined as having five or more drinks for men and four or more drinks for women on at least one occasion during the preceding month. Respondents who smoked  $\geq 100$  cigarettes in their lifetime and smoked every day or some days were classified as current smokers. Physical activity had following categories - meet recommendations for physical activity, insufficient physical activity, and do not participate in physical activity. Respondents who consumed 5 or more servings of fruits and vegetables everyday were classified as meet recommendation for fruit and vegetable consumption.

Chronic conditions or illness included were Cerebro-cardio-vascular disease, hypertension, had high cholesterol, diabetes, asthma, and overweight or obesity. Respondents were considered to have myocardial infarction, or angina, or stroke or high blood pressure if they had ever been told by a doctor, nurse, or other health professional to have myocardial infarction or stroke or high blood pressure respectively. Respondents were classified as having high blood cholesterol if they had checked their blood cholesterol and was told by a health professional that their blood cholesterol was high. Respondents were classified as having diabetes if they had ever been told by a doctor that they had diabetes. Asthma was self reported and physician or health care professional diagnosed; it had three categories - current asthma, former asthma, and never asthma. Self-reported weight and height were used to calculate Body Mass Index (BMI) ( $\text{BMI} = \text{weight}[\text{kg}] / (\text{height}[\text{m}])^2$ ). Participants were classified as overweight if their BMI was  $\geq 25 \text{ kg/m}^2$  and were classified as obese if their BMI was  $\geq 30 \text{ kg/m}^2$ .

We estimated the percent differences between early and late responders by demographics, health behaviors and chronic health conditions or illness. We used SUDAAN and

SAS for the analysis (SAS Institute Inc., Cary, NC, USA 2004).

### 3. Results

In the 2007 BRFSS survey, there were 430,912 interviews completed in the U.S. We excluded 14,189 records from two states (Michigan and Louisiana) and 49 cases with missing information. We analyzed the remaining 416,674 respondents of which 394,427 (95%) were early responders, and 22,247 (5%) were late responders. We estimated weighted and unweighted percent differences between early and late responders. The absolute differences between the weighted and unweighted percentages in the variables examined ranged between 0.06% and 2.6%, except white non-Hispanics where the absolute difference was 7%. We presented the unweighted analysis for the purpose of this study.

Significant differences were observed between early and late responders in demographics, access to health-care coverage, and general health status variables (Table 1). Compared to early responders, late responders were significantly more likely to be male, to report race/ethnicity as Hispanic, to have annual income of  $\geq \$50,000$ , to be younger than 45 years of age, to have less than high school education, to have access to health-care coverage, and to report good health. The absolute value of these significant differences in the variables above ranged from 1.3% to 7.6%. The percentage of Unknowns in the health-care coverage variable was 21% for late responders and 30% for early responders. The difference between early and late responders remained significant, even when we assumed the Unknowns to have a similar percentage of access to health-care coverage to those with known status in each respondent group.

A significant difference between early and late responders was also observed in health risk behaviors (Table 2). Compared to early responders, late responders were significantly less likely to meet the recommended guidelines for physical activity and daily consumption of fruits and vegetables. The absolute value of these significant differences ranged from 1.7 % to 3.1%. The differences between early and late responders remained significant even when the Unknowns were assumed to have a similar percentage to those of known status for both variables.

Table 3 shows the differences between early and late responders in chronic disease conditions or illnesses. Compared to early responders, late responders were significantly more likely to report high cholesterol, significantly less likely to report hypertension and diabetes, and were significantly less likely to be obese. The absolute value of these significant differences ranged from 1.8% to 5.8%.

**Table 1**  
**Percent differences between early responders and late responders by demographics, health-care coverage and general health, BRFSS 2007**

Demographics	Length of data collection		Difference (Early-late) %	P-Value
	Early responders* (N = 394,427) %	Late responders** (N = 22,247) %		
Gender				
Female	62.8	60.2	2.5	0.000
Male	37.3	39.8	-2.5	
Race				
White non-Hispanic	79.1	71.5	7.6	0.000
Black non-Hispanic	7.3	8.2	-0.9	0.168
Hispanic	7.1	13.5	-6.4	0.000
Others	5.5	5.8	-0.3	0.635
Unknown	1.0	1.0	0.0	0.977
Income				
<15,000	9.7	8.7	1.0	0.146
15-34,999	26.1	24.3	1.8	0.004
35-49,999	14.1	13.4	0.8	0.252
50,000+	36.6	39.7	-3.1	0.000
Unknown	13.5	14.0	-0.4	0.496
Age				
18-24	3.6	4.9	-1.3	0.025
25-44	25.7	33.3	-7.6	0.000
45-64	40.9	40.6	0.3	0.612
65+	29.0	20.2	8.8	0.000
Unknown	0.8	1.0	-0.1	0.827
Education Level				
<High School	10.3	12.3	-2.0	0.001
High School Graduate	30.6	28.7	1.9	0.001
> High School	58.8	58.2	0.6	0.177
Unknown	0.3	0.8	-0.5	0.264
Health care coverage (<65 years)				
Yes	59.3	65.4	-6.2	0.000
No	10.8	13.2	-2.5	
Unknown	30.0	21.4	8.6	
Health Status				
Good health	80.1	81.8	-1.7	0.000
Fair or poor health	19.4	17.6	1.8	
Unknown	0.5	0.6	-0.1	

\*Completed the survey within 31 days.

\*\*Completed the survey after 31 days.

**Table 2**  
**Percent differences between early responders and late responders by health risk behaviors, BRFSS 2007**

Risk factors	Length of data collection		Difference (Early-late) %	P-Value
	Early responders* (N = 394,427) %	Late responders** (N = 22,247) %		
Binge drinking				
Yes	11.1	11.8	-0.7	0.261
No	86.9	82.8	4.1	
Unknown	1.9	5.4	-3.4	
Smoking cigarettes				
Current smokers	18.3	17.5	0.9	0.182
Not a smoker	81.3	82.1	-0.8	
Unknown	0.4	0.5	0.0	
Physical activity recommendations				
Met recommended moderate/vigorous activity	43.4	41.8	1.7	0.000
Insufficient physical activity	35.4	31.8	3.6	
No physical activity	14.3	11.3	3.0	
Unknown	6.9	15.2	-8.3	
Fruit & vegetable consumption				
Consumed ≥ 5 times/day	25.0	21.9	3.1	0.000
Consumed < 5 times/day	73.0	69.7	3.3	
Unknown	2.0	8.5	-6.4	

\*Completed the survey within 31 days.

\*\*Completed the survey after 31 days.



**Table 3****Percent differences between early responders and late responders by chronic conditions and illnesses, BRFSS 2007**

Diseases/chronic conditions	Length of data collection		Difference (Early-late) %	P-Value
	Early responders* (N = 394,427) %	Late responders** (N = 22,247) %		
Cerebral and CVD:				
Myocardial Infarction				
Yes	5.9	4.9	1.0	0.177
No	93.6	94.7	-1.1	
Unknown	0.5	0.4	0.1	
Angina				
Yes	6.0	4.5	1.5	0.053
No	93.1	94.7	-1.6	
Unknown	0.9	0.8	0.1	
Stroke				
Yes	3.8	2.8	1.0	0.183
No	95.9	97.0	-1.1	
Unknown	0.3	0.2	0.1	
Other illnesses/conditions:				
High cholesterol				
Yes	57.0	60.8	-3.8	0.000
No	42.3	38.4	3.8	
Unknown	0.8	0.8	0.0	
Hypertension				
Yes	35.8	30.1	5.8	0.000
No	64.0	69.8	-5.8	
Unknown	0.2	0.2	0.0	
Diabetes				
Yes	11.2	9.4	1.8	0.010
Yes-Pregnancy	0.9	1.2	-0.2	
No	86.4	88.2	-1.9	
Borderline	1.4	1.2	0.2	
Unknown	0.1	0.1	0.0	
Asthma				
Current	8.7	7.7	1.0	0.158
Former	3.8	4.0	-0.2	
Never	86.9	87.8	-0.8	
Unknown	0.6	0.6	0.1	
Overweight or Obese				
Normal weight	34.5	35.5	-1.1	0.000
Over weight	35.0	34.7	0.4	
Obese	26.0	23.6	2.4	
Unknown	4.5	6.2	-1.7	

\*Completed the survey within 31 days.

\*\*Completed the survey after 31 days.

#### 4. Discussion

Our study found significant differences between early and late responders in demographic factors, and in some of the health risk behaviors and chronic disease conditions or illnesses. This shows that the composition of the two groups of responders is different with respect to these attributes. The differences observed could be due to difficulty in reaching persons working long hours and being away from their residences.

The greater likelihood of earning high income, being Hispanic, being young (18-44 years), having health-care coverage, having less than high school education, and reporting good general health among late responders fits the described characteristics of working people and healthy

workers (Li and Sung 1999), (O'Neil 1979). This description is supported by their significantly lower likelihood of reporting hypertension, diabetes and obesity. But certain risk behaviors show a different profile among late responders. Late responders are less likely to meet recommended guidelines for moderate or vigorous physical activity and for daily consumption of fruits and vegetables, which may be related to late responders having long working hours and poor access to healthy foods.

The high income earners, who are mostly white non-Hispanics, and low income earners, who are mostly Hispanics and black non-Hispanics, may spend long hours in their working environments and less likely to be in their homes to receive survey calls (Voigt, Koepsell and Daling 2003). In addition, BRFSS data indicate that interviewers

make more calls on late responders, on average almost 3 times more than on early responders, which bears out the difficulty of reaching them during the 31-day survey period. The reasons for working long hours could be different in the two income groups. Hispanics, black non-Hispanics, and young age groups may have low-paying jobs and need to work long hours to make a living, while the high-income individuals may have jobs requiring them to remain at work after regular working hours.

Surveillance and epidemiological estimates based only on early or late responders should be scrutinized for possible biases prior to making any generalizations. The percentage of interviews completed after 31 days is currently small (5%) and excluding them from the analysis may have no influence on national and state level estimates. However, as the proportions of late responders are expected to increase in the future, the influence of late responders on these estimates could not be ignored (Diehr, Cain, Connell and Volinn 1990). In addition, states should examine the consequences of ending data collection at 31 days on their operations, performance indicators, data quality measures, cost-savings and other contractual agreements with their data collection contractors.

Our study has a few limitations. BRFSS uses RDD methodology to select telephone numbers, which is subject to coverage bias (Rao, Link, Battaglia, Frankel, Giambo, and Mokdad 2005; Frankel, Srinath, Hoaglin, Battaglia, Smith, Wright and Khare 2003). Information collected is self-reported and may be subject to recall bias in some risk behaviors and disease estimations (Troiano, Berrigan, Dodd, Masse, Tilert and McDowell 2008; CDC 2004). In addition, we excluded two states from our analysis (Michigan and Louisiana), and extrapolation of the findings to these states should be done cautiously.

Despite these limitations, this study shows that late responders are significantly different in many respects from early responders. As the proportion of late responders may increase in the future, the influence of late responders on surveillance estimates should be examined carefully.

### Acknowledgements

We would like to acknowledge State BRFSS coordinators and their contractors.

### References

- Centers for Disease Control and Prevention (2009a). BRFSS Turning Information into Public Health, URL <http://www.cdc.gov/brfss/about.htm>.
- Centers for Disease Control and Prevention (2004). Access to health-care and preventive services among Hispanics and non-Hispanics – United States, 2001-2002. *Morbidity and Mortality Weekly Report*, 53, 937-941.
- Diehr, P., Cain, K., Connell, F. and Volinn, E. (1990). What is too much variation? The null hypothesis in small-area analysis. *Health Services Research*, 24, 741-771.
- Frankel, M.R., Srinath, K.P., Hoaglin, D.C., Battaglia, M.P., Smith, P.J., Wright, R.A. and Khare, M. (2003). Adjustments for non-telephone bias in random-digit-dialing surveys. *Statistics in Medicine*, 22, 1611-1626.
- Li, C.Y., and Sung, F.C. (1999). A review of the healthy worker effect in occupational epidemiology. *Occupational Medicine*, 49, 225-229.
- O'Neil, M.J. (1979). Estimating the nonresponse bias due to refusals in the telephone surveys. *Public Opinion Quarterly*, 43, 218-232.
- Qayad, M.G., Balluz, L. and Garvin, W. (2009). Does continuing data collection beyond one month improve the completion and response rates in behavioral risk factor surveillance system survey? *Survey Practice Feb 2009*. URL <http://surveypractice.org/2009/02/>.
- Rao, R.S., Link, M.W., Battaglia, M.P., Frankel, M.R., Giambo, P. and Mokdad, A.H. (2005). Assessing representativeness in RDD surveys: coverage and non-response in the Behavioral Risk Factor Surveillance System. Minneapolis: Joint Statistical Meetings. URL <http://www.amstat.org/sections/SRMS/Proceedings/y2005/Files/JSM2005-000190.pdf>.
- SAS Institute Inc. (2004). SAS version 9.1, Cary, NC: SAS Institute, USA.
- Troiano, R.P., Berrigan, D., Dodd, K.W., Masse, L.C., Tilert, T. and McDowell, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise*, 40, 181-8.
- Voigt, L.F., Koepsell, T.D. and Daling, J.R. (2003). Characteristics of Telephone Survey Respondents According to Willingness to Participate. *American Journal of Epidemiology*, 157, 66-73.





# An interesting property of the entropy of some sampling designs

Yves Tillé and David Haziza<sup>1</sup>

## Abstract

In this short note, we show that simple random sampling without replacement and Bernoulli sampling have approximately the same entropy when the population size is large. An empirical example is given as an illustration.

Key Words: Conditional Poisson sampling; Entropy; Simple random sampling; Poisson sampling.

## 1. Introduction

Consider a finite population of size  $N$  and let  $U = \{1, \dots, k, \dots, N\}$  be the set of labels of this population. A sample  $s$  is a subset of  $U$  and a sampling design is a probability law  $p(\cdot)$  on the subsets of  $U$  such that  $p(s) \geq 0$  for all  $s \subset U$ , and

$$\sum_{s \subset U} p(s) = 1.$$

Let  $\pi_k = P(k \in s)$  be the first-order inclusion probability of unit  $k$  in the sample:

$$\pi_k = \sum_{\substack{s \subset U \\ s \ni k}} p(s).$$

Similarly, let  $\pi_{k\ell} = P(k \in s \text{ and } \ell \in s)$  be the second-order inclusion probability of unit  $k$  and  $\ell$  in the sample:

$$\pi_{k\ell} = \sum_{\substack{s \subset U \\ s \ni k, \ell}} p(s).$$

The entropy of a sampling design  $p(\cdot)$ , denoted by  $I(p)$ , is defined as

$$I(p) = - \sum_{s \in Q} p(s) \log p(s), \quad (1)$$

where  $Q = \{s | p(s) > 0\}$  is the support of the sampling design  $p(\cdot)$ . A sampling design has high entropy when there is a high amount of uncertainty or high amount of surprise in the sample which will be selected. In other words, when a sampling design has high entropy, it is very difficult to predict the type of sample we would obtain. Many sampling designs used in practice are high entropy designs. One notable exception is systematic sampling that has a very low entropy. The concept of entropy is useful in the context of variance estimation. When a sampling design has a high entropy, it is possible to obtain approximation of the second-order inclusion probabilities,  $\pi_{k\ell}$ , in terms of the first-order inclusion probabilities, which simplifies considerably the problem of variance estimation in the

context of unequal probability sampling; *e.g.*, Brewer and Donadio (2003), Matei and Tillé (2005), Henderson (2006) and Haziza, Mecatti and Rao (2008).

It is well known that the sampling design with maximum entropy is Poisson sampling:

$$p_{\text{poiss}}(s) = \left( \prod_{k \in s} \pi_k \right) \left( \prod_{k \in U \setminus s} (1 - \pi_k) \right) \quad (2)$$

for all  $s \in Q$ ; *e.g.*, Tillé (2006). A special case of Poisson sampling is Bernoulli sampling, which is obtained from (2) by setting  $\pi_k = \pi \in (0, 1)$ , which leads to

$$p_{\text{bern}}(s) = \pi^{n_s} (1 - \pi)^{N - n_s}, \text{ for all } s \subset U,$$

where  $n_s$  is the random size of  $s$ . Using (1) and noting that  $\sum_{s \in Q} n_s p(s) = N\pi$ , the entropy of Bernoulli sampling is given by

$$I(p_{\text{bern}}) = -N(1 - \pi) \log(1 - \pi) - N\pi \log \pi, \quad (3)$$

which is maximum when  $\pi = 1/2$ . In this case, we have  $I(p_{\text{bern}}) = N \log 2$ .

If we restrict to the class of fixed size sampling designs with first-order inclusion probabilities  $\pi_k$ ,  $k \in U$ , the maximum entropy design is the so-called Conditional Poisson Sampling (CPS); (see Chen, Dempster and Liu 1994; Deville 2000; Tillé 2006). The CPS design can be implemented by repeatedly selecting samples according to Poisson sampling until the desired sample size,  $n$  (say), has been obtained. When  $\pi_k = n/N$  for all  $k \in U$ , the CPS design reduces to simple random sampling without replacement:

$$p_{\text{srs}}(s) = \binom{N}{n}^{-1}$$

for all  $s \in Q$ . From (1), it follows that the entropy of simple random sampling is given by

$$I(p_{\text{srs}}) = \log N! - \log n! - \log(N - n)!. \quad (4)$$

In other words, simple random sampling without replacement is the maximum entropy design in the class of equal probability fixed size sampling designs.

1. Yves Tillé, Institut de Statistique, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland; David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, QC, Canada, H3C 3J7. E-mail: haziza@dms.umontreal.ca.

Not all sampling designs possess a high entropy. For example, the 1-in- $G$  systematic sampling design has a very low entropy. Here, the number of samples,  $G = N/n$ , is assumed to be an integer value. Since  $p_{\text{syst}}(s) = 1/G$  for all  $s \in Q$ , the entropy of systematic sampling is given by

$$I(p_{\text{syst}}) = \log N - \log n,$$

which is much smaller than (4), especially for large values of  $N$ .

2. Main result

In this section, we compare the entropy of Bernoulli sampling with that of simple random sampling without replacement. Since the support of the Bernoulli sampling designs is much larger than that of simple random sampling without replacement, we expected the entropy of Bernoulli sampling to be much larger than that of simple random sampling without replacement. Table 1 shows the entropy for simple random sampling and Bernoulli sampling for different values of  $N$  and  $\pi$ . Surprisingly, we found the entropy of both sampling designs for the same inclusion probabilities and the same sample size to be approximately equal. From Table 1, it is clear that both sampling designs have similar entropies, even for moderate population sizes (e.g.,  $N = 100$ ), independently of the value of  $\pi$ . This result is somehow curious considering the strong reduction of possible samples by fixing the sample size. Indeed, recall that the size of the support is  $\binom{N}{n}$  for simple random sampling without replacement, whereas it is  $2^N$  for Bernoulli sampling. For example, for  $N = 100$  and  $n = 20$ , the size of the support for simple random sampling without replacement is equal to  $\binom{100}{20} \approx 5.36 \times 10^{20}$ , whereas it is equal to  $2^{100} \approx 1.26 \times 10^{30}$  for Bernoulli sampling. In other words, the size of the support of Bernoulli sampling is approximately  $2.36 \times 10^9$  larger than that of simple random sampling without replacement.

*Result 1.* Let  $I(p_{\text{bern}})$  and  $I(p_{\text{srs}})$  be the entropy for Bernoulli sampling and simple random sampling without replacement, respectively given by (3) and (4). Then,

$$\lim_{N \rightarrow \infty} \frac{I(p_{\text{srs}})}{I(p_{\text{bern}})} = 1.$$

*Proof.* By considering Stirling’s formula (see Abramowitz and Stegun 1964, page 257)

$$\lim_{n \rightarrow \infty} \frac{n \log n - n}{\log n!} = 1,$$

we get

$$\lim_{\substack{N \rightarrow \infty \\ \pi \rightarrow \infty \\ N-n \rightarrow \infty}} \frac{N \log N - n \log n - (N - n) \log (N - n)}{\log \binom{N}{n}} = 1,$$

from which we obtain

$$\lim_{N \rightarrow \infty} \frac{\log \binom{N}{N\pi}}{-N(1 - \pi) \log (1 - \pi) - N\pi \log \pi} = 1.$$

3. Conclusion

In this note, we showed that Bernoulli sampling and simple random sampling without replacement have very similar entropies, even for moderate population sizes. We conjecture that the same should be observed when comparing the Poisson sampling design and the CPS design for a given set on first-order inclusion probabilities. However, the proof of this result seems to be considerably more complex.

Table 1  
Entropy of (Bernoulli sampling, simple random sampling) designs

$N$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$	$\pi = 0.4$	$\pi = 0.5$
10	(3.3, 2.3)	(5, 3.8)	(6.1, 4.8)	(6.7, 5.3)	(6.9, 5.5)
100	(32.5, 30.5)	(50, 47.7)	(61.1, 58.6)	(67.3, 64.8)	(69.3, 66.8)
1,000	(325.1, 321.9)	(500.4, 496.9)	(610.9, 607.3)	(673, 669.4)	(693.1, 689.5)
10,000	(3,250.8, 3,246.5)	(5,004, 4,999.4)	(6,108.6, 6,103.9)	(6,730.1, 6,725.3)	(6,931.5, 6,926.6)
100,000	(32,508.3, 32,502.8)	(50,040.2, 50,034.5)	(61,086.4, 61,080.5)	(67,301.2, 67,295.2)	(69,314.7, 69,308.7)
1,000,000	(325,083, 325,076)	(500,402, 500,396)	(610,864, 610,857)	(673,012, 673,005)	(693,147, 693,140)

### Acknowledgements

We thank an Associate Editor and a referee for constructive comments. Work of David Haziza was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

### References

- Abramowitz, M., and Stegun, I.A. (1964). *Handbook of Mathematical Functions*. New York: Dover.
- Brewer, K.R.W., and Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29, 189-196.
- Chen, S.X., Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.
- Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI, Rennes.
- Haziza, D., Mecatti, F. and Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron*, 66, 91-108.
- Henderson, T. (2006). Estimating the variance of the Horvitz-Thompson estimator. Master's thesis, School of Finance and Applied Statistics, The Australian National University.
- Matei, A., and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21, 4, 543-570.
- Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.





## ACKNOWLEDGEMENTS

*Survey Methodology* wishes to thank the following people who have provided help or served as referees for one or more papers during 2010.

- P. Ardilly, *INSEE*  
P. Beatty, *National Center for Health Statistics*  
J.-F. Beaumont, *Statistics Canada*  
C. Bocci, *Statistics Canada*  
G. Brackstone  
J. van den Brakel, *Statistics Netherlands*  
J.M. Brick, *Westat Inc*  
P.D. Brick, *Westat Inc*  
C. Calder, *Ohio State University*  
P. Cantwell, *U.S. Bureau of the Census*  
R. Chambers, *Centre for Statistical and Survey Methodology*  
P. Dick, *Statistics Canada*  
A.H. Dorfman, *U.S. Bureau of Labour Statistics*  
G. Dubreuil, *Statistics Canada*  
M. Elliott, *University of Michigan*  
J.L. Eltinge, *U.S. Bureau of Labor Statistics*  
G. Falk, *University of Virginia*  
M. Fay, *National Institute of Allergy and Infectious Diseases*  
W.A. Fuller, *Iowa State University*  
J. Gambino, *Statistics Canada*  
N. Ganesh, *National Opinion Research Center at the University of Chicago*  
S. Ghosh, *Alberta Health Services-Cancer care*  
C. Girard, *Statistics Canada*  
S. Godbout, *Statistics Canada*  
R. Griffin, *U.S. Census Bureau*  
D. Haziza, *Université de Montréal*  
Y. He, *Harvard Medical School*  
S. Heeringa, *University of Michigan*  
M. Hidirolou, *Statistics Canada*  
S. Holan, *University of Missouri*  
A. Holmberg, *Statistics Sweden*  
D. Hubble, *Westat Inc*  
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*  
J. Jones, *Office for National Statistics, United Kingdom*  
D. Judkins, *Westat Inc*  
D. Kasprzyk, *Mathematica Policy Research*  
P. Kelly, *Statistics Canada*  
P. Kott, *RTI International*  
P. Lahiri, *JPSM, University of Maryland*  
M.D. Larsen, *George Washington University*  
P. Lavallée, *Statistics Canada*  
J. Legg, *Amgen Inc., USA*  
R. Little, *University of Michigan*  
S. Lohr, *Arizona State University*  
P. Lynn, *University of Essex*  
M. Maia, *Catholic University of Portugal*  
D.J. Malec, *U.S. Census Bureau*  
J. Maples, *U.S. Census Bureau*  
E. Martin, *U.S. Census Bureau*  
K. Miller, *U.S. National Center for Health Statistics*  
T. Mulcahy, *National Opinion Research Center*  
G. Nathan, *Hebrew University*  
S.F. Nielsen, *Copenhagen Business School, Denmark*  
A. Nigam, *Institute of Applied Statistics and Development Studies*  
J. Opsomer, *Colorado State University*  
Z. Patak, *Statistics Canada*  
D. Pfeffermann, *Hebrew University*  
N.G.N. Prasad, *University of Alberta*  
M. Pratesi, *Università di Pisa*  
J.N.K. Rao, *Carleton University*  
J. Reiter, *Duke University*  
L.-P. Rivest, *Université Laval*  
S. Rubin-Bleuer, *Statistics Canada*  
J. Ryten  
N. Salvati, *Università di Pisa*  
C.-E. Särndal, *Université de Montréal*  
N. Schenker, *National Center for Health Statistics*  
F.J. Scheuren, *National Opinion Research Center*  
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*  
P. Smith, *Office for National Statistics*  
T.M.F. Smith, *University of Southampton, UK*  
E. Stasny, *Ohio State University*  
D. Steel, *University of Wollongong*  
L. Stokes, *Southern Methodist University*  
A. Théberge, *Statistics Canada*  
M. Thompson, *University of Waterloo*  
S. Thompson, *Simon Fraser University*  
D. Toth, *U.S. Bureau of Labor Statistics*  
C. Tucker, *U.S. Bureau of Labor Statistics*  
V.J. Verma, *Università degli Studi di Siena*  
W.E. Winkler, *U.S. Census Bureau*  
K.M. Wolter, *Iowa State University*  
C. Wu, *University of Waterloo*  
W. Yung, *Statistics Canada*  
P.A. Zandbergen, *University of New Mexico*  
A. Zaslavsky, *Harvard Medical School*

Acknowledgements are also due to those who assisted during the production of the 2010 issues: Céline Ethier of Statistical Research and Innovation Division, Christine Cousineau and Teresa Jewell of Household Survey Methods Division, Nick Budko and Sophie Chartier of Business Survey Methods Division, Matthew Belyea, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Darquise Pellerin of Dissemination Division, and Jeff Jodoin of Client Services Division.





## ANNOUNCEMENTS

### Nominations Sought for the 2012 Waksberg Award

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg to recognize his contributions to survey methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium from Westat. The paper will be published in a future issue of *Survey Methodology*.

The author of the 2012 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nomination of individuals to be considered as authors or suggestions for topics should be sent before February 28, 2011 to the chair of the committee, Elizabeth Martin ([betsy@folhc.org](mailto:betsy@folhc.org)).

Previous Waksberg Award honorees and their invited papers are:

- 2001 Gad **Nathan**, "Telesurvey methodologies for household surveys – A review and some thoughts for the future?". *Survey Methodology*, vol. 27, 1, 7-31.
- 2002 Wayne A. **Fuller**, "Regression estimation for survey samples". *Survey Methodology*, vol. 28, 1, 5-23.
- 2003 David **Holt**, "Methodological issues in the development and use of statistical indicators for international comparisons". *Survey Methodology*, vol. 29, 1, 5-17.
- 2004 Norman M. **Bradburn**, "Understanding the question-answer process". *Survey Methodology*, vol. 30, 1, 5-15.
- 2005 J.N.K. **Rao**, "Interplay between sample survey theory and practice: An appraisal". *Survey Methodology*, vol. 31, 2, 117-138.
- 2006 Alastair **Scott**, "Population-based case control studies". *Survey Methodology*, vol. 32, 2, 123-132.
- 2007 Carl-Erik **Särndal**, "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.
- 2008 Mary E. **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.
- 2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.
- 2010 Ivan P. **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.
- 2011 Danny **Pfeffermann**, Manuscript topic under consideration.

**Members of the Waksberg Paper Selection Committee (2010-2011)**

Elizabeth A. Martin (Chair)  
Mary Thompson, *University of Waterloo*  
J.N.K. Rao, *Carleton University*  
Steve Heeringa, *University of Michigan*

**Past Chairs:**

Graham Kalton (1999 - 2001)  
Chris Skinner (2001 - 2002)  
David A. Binder (2002 - 2003)  
J. Michael Brick (2003 - 2004)  
David R. Bellhouse (2004 - 2005)  
Gordon Brackstone (2005 - 2006)  
Sharon Lohr (2006 - 2007)  
Robert Groves (2007 - 2008)  
Leyla Mojadjer (2008 - 2009)  
Daniel Kasprzyk (2009 - 2010)

## Volume 38, No. 2, June/juin 2010

**Special Issue: Inferences in Generalized Linear Longitudinal Mixed Models**

Brajendra Sutradhar	
Preface to the special issue on inferences in generalized linear longitudinal mixed models.....	171
Brajendra C. Sutradhar	
Inferences in generalized linear longitudinal mixed models.....	174
Marco Alfò and Antonello Maruotti	
Two-part regression models for longitudinal zero-inflated count data .....	197
Brajendra C. Sutradhar and Taslim S. Mallick	
Modified weights based generalized quasiliikelihood inferences in incomplete longitudinal binary models.....	217
Grace Y. Yi, Richard J. Cook and Baojiang Chen	
Estimating functions for evaluating treatment effects in cluster-randomized longitudinal studies in the presence of drop-out and non-compliance .....	232
Josue G. Martinez, Faming Liang, Lan Zhou and Raymond J. Carroll	
Longitudinal functional principal component modelling via Stochastic Approximation Monte Carlo.....	256
Emily L. Kang, Noel Cressie and Tao Shi	
Using temporal variability to improve spatial mapping with application to satellite data .....	271
Brajendra Sutradhar, Alwell J. Oyet and Veeresh G. Gadag	
On quasi-likelihood estimation for branching processes with immigration.....	290

## Volume 38, No. 3, September/septembre 2010

Hongying Dai and Richard Charnigo	
Contaminated normal modeling with application to microarray data analysis .....	315
Xinyuan Song, Liuquan Sun, Xiaoyun Mu and Gregg E. Dinse	
Additive hazards regression with censoring indicators missing at random.....	333
Bo Hu, Jun Shao and Mari Palta	
Variability explained by covariates in linear mixed-effect models for longitudinal data.....	352
Isabel Molina and J.N.K. Rao	
Small area estimation of poverty indicators.....	369
Grace S. Chiu and Richard A. Lockhart	
Bent-cable regression with autoregressive noise .....	386
Haiyan Wang, Siti Tolos and Suojin Wang	
A distribution free test to detect general dependence between a response variable and a covariate in the presence of heteroscedastic treatment effects.....	408
Zhensheng Huang and Riquan Zhang	
Empirical likelihood for the varying-coefficient single-index model .....	434
Elisa M. Molanes-lopez, Ricardo Cao and Ingrid VAN Keilegom	
Smoothed empirical likelihood confidence intervals for the relative distribution with left-truncated and right-censored data .....	453
Robert Paige and Edward Allen	
Closed-form likelihoods for stochastic differential equation growth models .....	474
Terry C.K. Lee, Min Tsao and Francis W. Zwiers	
State-space model for proxy-based millennial reconstruction .....	488
Hae-Ryoung Song, Andrew B. Lawson and Daniela Nitcheva	
Bayesian hierarchical models for food frequency assessment .....	506

## ERRATA

Pierre Duchesne and Simon Lalancette	
Erratum: Authors' corrigenda/corrections des auteurs on testing for multivariate ARCH effects in vector time series models....	517
Article first published online: 31 AUG 2010   DOI: 10.1002/cjs.10067	
<b>This article corrects:</b>	
On testing for multivariate ARCH effects in vector time series models	
Vol. 31, Issue 3, 275–292, Article first published online: 18 DEC 2008	



# JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents

Volume 26, No. 1, 2010

Special Section with Articles Based on Papers from the Third International Conference on Establishment Surveys – Preface.....	1
A Hybrid Response Process Model for Business Surveys Diane K. Willimack, Elizabeth Nichols.....	3
Sources of Measurement Errors in Business Surveys Mojca Bavdaž.....	25
Questionnaire Design Guidelines for Establishment Surveys Rebecca L. Morrison, Don A. Dillman, Leah M. Christian.....	43
From Start to Pilot: A Multi-method Approach to the Comprehensive Redesign of an Economic Survey Questionnaire Alfred D. Tuttle, Rebecca L. Morrison, Diane K. Willimack.....	87
Adjusting for Nonignorable Sample Attrition Using Survey Substitutes Identified by Propensity Score Matching: An Empirical Investigation Using Labour Market Data Richard Dorsett.....	105
Evaluation and Selection of Models for Attrition Nonresponse Adjustment Eric V. Slud, Leroy Bailey.....	127
Trends in Income Nonresponse Over Two Decades Ting Yan, Richard Curtin, Matthew Jans.....	145
Get It or Drop It? Cost-Benefit Analysis of Attempts to Interview in Household Surveys Dmitri Romanov, Michal Nir.....	165
Comparing Four Bootstrap Methods for Stratified Three-Stage Sampling Hiroshi Saigo.....	193
Book and Software Review.....	209
In Other Journals.....	213

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)

# JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents

Volume 26, No. 2, 2010

Official Statistics in India: The Past and the Present T.J. Rao .....	215
Changing from PAPI to CAPI: Introducing CAPI in a Longitudinal Study Jörg-Peter Schräpler, Jürgen Schupp, Gert G. Wagner .....	233
Effects of Answer Space Size on Responses to Open-ended Questions in Mail Surveys Glenn D. Israel.....	271
Experimental Studies of Disclosure Risk, Disclosure Harm, Topic Sensitivity, and Survey Participation Mick P. Couper, Eleanor Singer, Frederick G. Conrad, Robert M. Groves.....	287
Tests of Multivariate Hypotheses when using Multiple Imputation for Missing Data and Disclosure Limitation Satkartar K. Kinney, Jerome P. Reiter.....	301
Issues in Survey Measurement of Chronic Disability: An Example from the National Long Term Care Survey Elena A. Erosheva, Toby A. White.....	317
District-level Estimates of Institutional Births in Ghana: Application of Small Area Estimation Technique Using Census and DHS Data Fiifi Amoako Johnson, Hukum Chandra, James J. Brown, Sabu S. Padmadas .....	341
Seasonality in Revisions of Macroeconomic Data Philip Hans Franses, Rene Segers .....	361
Comparison of X-12-ARIMA Trading Day and Holiday Regressors with Country Specific Regressors Christopher G. Roberts, Scott H. Holan, Brian Monsell .....	371

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)

# JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents Volume 26, No. 3, 2010

The 2009 Morris Hansen Lecture: The Care, Feeding, and Training of Survey Statisticians Sharon L. Lohr.....	395
Discussion	
Donsig Jang .....	411
James M. Lepkowski .....	417
David Morganstein.....	421
The Role of the Joint Program in Survey Methodology in Training U.S. Federal Statisticians Richard Valliant, Roger Tourangeau, Janice Lent.....	427
Statistical Careers in United States Government Science Agencies Nell Sedransk.....	443
Recruitment, Training and Retention of Statisticians in the U.S. Federal Statistical Agencies Clyde Tucker .....	455
Contact Strategies to Improve Participation via the Web in a Mixed-Mode Mail and Web Survey Anders Holmberg, Boris Lorenc, Peter Werner.....	465
Comparison of Verbal Behaviors between Calendar and Standardized Conventional Questionnaires Ipek Bilgen, Robert F. Belli.....	481
Using Audio Computer-Assisted Self-Interviewing and Interactive Voice Response to Measure Elder Mistreatment in Older Adults: Feasibility and Effects on Prevalence Estimates Scott R. Beach, Richard Schulz, Howard B. Degenholtz, Nicholas G. Castle, Jules Rosen, Andrea R. Fox, Richard K. Morycz.....	507
Bilingual Questionnaire Evaluation and Development through Mixed Pretesting Methods: The Case of the U.S. Census Nonresponse Followup Instrument Jennifer Childs, Patricia Goerman .....	535
Using XBRL in a Statistical Context. The Case of the Dutch Taxonomy Project Marko Roos .....	559
Book Reviews .....	577
In Other Journals .....	583

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)



# GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A pdf or paper copy may be required for formulas and figures.

## 1. Layout

- 1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The documents should be divided into numbered sections with suitable verbal titles.
- 1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

## 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

## 3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis.

## 4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

## 5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

## 6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

# DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de finaliser votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1. **Présentation**
  - 1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
  - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
  - 1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
  - 1.4 Les remerciements doivent paraître à la fin du texte.
  - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
  - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
  - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme  $\exp(\cdot)$  et  $\log(\cdot)$  etc.
  - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
  - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
  - 3.5 Distinguer clairement les caractères ambigus (comme w,  $\omega$  ; o, O, 0 ; l, 1).
  - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
4. **Figures et tableaux**

Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

5. **Bibliographie**
  - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).
  - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
6. **Communications brèves**

Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.







JOURNAL OF OFFICIAL STATISTICS  
An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 26, No. 3, 2010

The 2009 Morris Hansen Lecture: The Care, Feeding, and Training of Survey Statisticians  
Sharon L. Lohr..... 395

Discussion  
Donsig Jang ..... 411  
James M. Lepkowski ..... 417  
David Morganstein..... 421

The Role of the Joint Program in Survey Methodology in Training U.S. Federal Statisticians  
Richard Valliant, Roger Tourangeau, Janice Lent..... 427

Statistical Careers in United States Government Science Agencies  
Neil Sedransk..... 443

Recruitment, Training and Retention of Statisticians in the U.S. Federal Statistical Agencies  
Clyde Tucker..... 455

Contact Strategies to Improve Participation via the Web in a Mixed-Mode Mail and Web Survey  
Anders Holmberg, Boris Lorenc, Peter Werner..... 465

Comparison of Verbal Behaviors between Calendar and Standardized Conventional Questionnaires  
Ipek Bilgen, Robert F. Belli..... 481

Using Audio Computer-Assisted Self-Interviewing and Interactive Voice Response to Measure Elder Mistreatment in Older Adults: Feasibility and Effects on Prevalence Estimates  
Scott R. Beach, Richard Schulz, Howard B. Degenholtz, Nicholas G. Castle, Jules Rosen, Andrea R. Fox, Richard K. Morycz..... 507

Bilingual Questionnaire Evaluation and Development through Mixed Pretesting Methods: The Case of the U.S. Census Nonresponse Followup Instrument  
Jennifer Childs, Patricia Goerman..... 535

Using XBRL in a Statistical Context. The Case of the Dutch Taxonomy Project  
Marko Roos..... 559

Book Reviews ..... 577

In Other Journals..... 583

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)

## JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents

Volume 26, No. 2, 2010

Official Statistics in India: The Past and the Present	T. J. Rao .....	215
Changing from PAPI to CAPI: Introducing CAPI in a Longitudinal Study	Jörg-Peter Schräpler, Jürgen Schupp, Gert G. Wagner .....	233
Effects of Answer Space Size on Responses to Open-ended Questions in Mail Surveys	Glenn D. Israel .....	271
Experimental Studies of Disclosure Risk, Disclosure Harm, Topic Sensitivity, and Survey Participation	Mick P. Couper, Eleanor Singer, Frederick G. Conrad, Robert M. Groves .....	287
Tests of Multivariate Hypotheses when using Multiple Imputation for Missing Data and Disclosure Limitation	Satkar K. Kinney, Jerome P. Reiter .....	301
Issues in Survey Measurement of Chronic Disability: An Example from the National Long Term Care Survey	Elena A. Erosheva, Toby A. White .....	317
District-level Estimates of Institutional Births in Ghana: Application of Small Area Estimation Technique Using Census and DHS Data	Fifti Amoko Johnson, Hukum Chandra, James J. Brown, Sabu S. Padmadas .....	341
Seasonality in Revisions of Macroeconomic Data	Philip Hans Franses, Rene Segers .....	361
Comparison of X-12-ARIMA Trading Day and Holiday Regressors with Country Specific Regressors	Christopher G. Roberts, Scott H. Holan, Brian Monsell .....	371

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)



# JOURNAL OF OFFICIAL STATISTICS

## An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

### Contents

#### Volume 26, No. 1, 2010

Special Section with Articles Based on Papers from the Third International Conference on Establishment Surveys – Preface.....	1
A Hybrid Response Process Model for Business Surveys Diane K. Willimack, Elizabeth Nichols.....	3
Sources of Measurement Errors in Business Surveys Mojca Bavdaz.....	25
Questionnaire Design Guidelines for Establishment Surveys Rebecca L. Morrison, Don A. Dillman, Leah M. Christian.....	43
From Start to Pilot: A Multi-method Approach to the Comprehensive Redesign of an Economic Survey Questionnaire Alfred D. Tuttle, Rebecca L. Morrison, Diane K. Willimack.....	87
Adjusting for Nonignorable Sample Attrition Using Survey Substitutes Identified by Propensity Score Matching: An Empirical Investigation Using Labour Market Data Richard Dorsett.....	105
Evaluation and Selection of Models for Attrition Nonresponse Adjustment Eric V. Slud, Leroy Bailey.....	127
Trends in Income Nonresponse Over Two Decades Ting Yan, Richard Curtin, Mathew Jans.....	145
Get It or Drop It? Cost-Benefit Analysis of Attempts to Interview in Household Surveys Dmitri Romanov, Michal Nir.....	165
Comparing Four Bootstrap Methods for Stratified Three-Stage Sampling Hiroshi Saigo.....	193
Book and Software Review.....	209
In Other Journals.....	213

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)

## CONTENTS

## TABLE DES MATIÈRES

## Volume 38, No. 2, June/juin 2010

## Special Issue: Inferences in Generalized Linear Longitudinal Mixed Models

Brajendra Sutradhar	Preface to the special issue on inferences in generalized linear longitudinal mixed models.....	171
Brajendra C. Sutradhar	Inferences in generalized linear longitudinal mixed models.....	174
Marco Alfio and Antonello Maruotti	Two-part regression models for longitudinal zero-inflated count data.....	197
Brajendra C. Sutradhar and Taslim S. Mallik	Modified weights based generalized quasiliikelihood inferences in incomplete longitudinal binary models.....	217
Grace Y. Yi, Richard J. Cook and Baofiang Chen	Estimating functions for evaluating treatment effects in cluster-randomized longitudinal studies in the presence of drop-out and non-compliance.....	232
Josue G. Martinez, Faming Liang, Lan Zhou and Raymond J. Carroll	Longitudinal functional principal component modelling via Stochastic Approximation Monte Carlo.....	256
Emily L. Kang, Noel Cressie and Tao Shi	Using temporal variability to improve spatial mapping with application to satellite data.....	271
Brajendra Sutradhar, Alwell J. Oyot and Veeresh G. Gadag	On quasi-likelihood estimation for branching processes with immigration.....	290

## Volume 38, No. 3, September/septembre 2010

Hongying Dai and Richard Chamigo	Contaminated normal modeling with application to microarray data analysis.....	315
Xinyuan Song, Linquan Sun, Xiaoyun Mu and Gregg E. Dinse	Additive hazards regression with censoring indicators missing at random.....	333
Bo Hu, Jun Shao and Mari Palta	Variability explained by covariates in linear mixed-effect models for longitudinal data.....	352
Isabel Molina and J.N.K. Rao	Small area estimation of poverty indicators.....	369
Grace S. Chiu and Richard A. Lockhart	Bent-cable regression with autoregressive noise.....	386
Haiyan Wang, Siti Tolos and Suojin Wang	A distribution free test to detect general dependence between a response variable and a covariate in the presence of heteroscedastic treatment effects.....	408
Zhensheng Huang and Riquan Zhang	Empirical likelihood for the varying-coefficient single-index model.....	434
Elisa M. Molanes-lopez, Ricardo Cao and Ingrid VAN Keilegom	Smoothed empirical likelihood confidence intervals for the relative distribution with left-truncated and right-censored data.....	453
Robert Paige and Edward Allen	Closed-form likelihoods for stochastic differential equation growth models.....	474
Terry C.K. Lee, Min Tsao and Francis W. Zwieters	State-space model for proxy-based millennial reconstruction.....	488
Hae-Ryoung Song, Andrew B. Lawson and Daniela Nitcheva	Bayesian hierarchical models for food frequency assessment.....	506
ERRATA		

Pierre Duchesne and Simon Lalancette

Erratum: Authors' corrigenda/corrections des auteurs on testing for multivariate ARCH effects in vector time series models.... 517  
 Article first published online: 31 AUG 2010 | DOI: 10.1002/cjs.10067  
 This article corrects:  
 On testing for multivariate ARCH effects in vector time series models  
 Vol. 31, Issue 3, 275–292, Article first published online: 18 DEC 2008

# Membres du comité de sélection de l'article Waksberg (2010-2011)

Elizabeth A. Martin (Présidente)  
 Mary Thompson, *University of Waterloo*  
 J.N.K. Rao, *Carleton University*  
 Steve Heeringa, *University of Michigan*

## Présidents précédents :

Graham Kalton (1999 - 2001)  
 Chris Skinner (2001 - 2002)  
 David A. Binder (2002 - 2003)  
 J. Michael Brick (2003 - 2004)  
 David R. Bellhouse (2004 - 2005)  
 Gordon Brackstone (2005 - 2006)  
 Sharon Lohr (2006 - 2007)  
 Robert Groves (2007 - 2008)  
 Leyla Mojadjer (2008 - 2009)  
 Daniel Kasprzyk (2009 - 2010)



## ANNONCES

## Demande de candidatures pour le prix Waksberg 2012

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'agencement de théorie et de pratique caractéristique des travaux de Waksberg.

L'auteur reçoit une prime en argent qui provient d'une bourse de Westat. L'article sera publié dans un numéro futur de *Techniques d'enquête*.

L'auteur de l'article Waksberg de 2012 sera sélectionné par un comité de quatre personnes désignées par *Techniques d'enquête* et l'American Statistical Association. Les candidatures ou les suggestions de sujets doivent être envoyées avant le 28 février 2010 à la présidente du comité Elizabeth Martin ([betsy@folhc.org](mailto:betsy@folhc.org)).

Les gagnants et articles précédents du prix Waksberg sont

- 2001 Gad Nathan, « Méthodes de télé-enquêtes applicables aux enquêtes-ménages – Revue et réflexions sur l'avenir », *Techniques d'enquête*, vol. 27, 1, 7-34.
- 2002 Wayne A. Fuller, « Estimation par régression appliquée à l'échantillonnage », *Techniques d'enquête*, vol. 28, 1, 5-25.
- 2003 David Holt, « Enjeux méthodologiques de l'élaboration et de l'utilisation d'indicateurs statistiques pour des fins de comparaisons internationales », *Techniques d'enquête*, vol. 29, 1, 5-19.
- 2004 Norman M. Bradburn, « Comprendre le processus de question et réponse », *Techniques d'enquête*, vol. 30, 1, 5-16.
- 2005 J.N.K. Rao, « Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage », *Techniques d'enquête*, vol. 31, 2, 127-151.
- 2006 Alastair Scott, « Études cas-témoins basées sur la population », *Techniques d'enquête*, vol. 32, 2, 137-147.
- 2007 Carl-Erik Särndal, « La méthode de calage dans la théorie et la pratique des enquêtes », *Techniques d'enquête*, vol. 33, 2, 113-135.
- 2008 Mary E. Thompson, « Enquêtes internationales : motifs et méthodologies », *Techniques d'enquête*, vol. 34, 2, 145-157.
- 2009 Graham Kalton, « Méthodes de surechantillonnage des sous-populations rares dans les enquêtes sociales », *Techniques d'enquête*, vol. 35, 2, 133-152.
- 2010 Ivan P. Fellegi, « L'organisation de la méthodologie statistique et de la recherche méthodologique dans les bureaux nationaux de la statistique », *Techniques d'enquête*, vol. 36, 2, 131-139.
- 2011 Danny Pfeffermann, Sujet de l'article à l'étude.



## REMERCIEMENTS

*Techniques d'enquête* désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2010.

- P. Ardilly, *INSEE*  
 P. Beatty, *National Center for Health Statistics*  
 J.-F. Beaumont, *Statistique Canada*  
 C. Bocci, *Statistique Canada*  
 G. Brackstone  
 J. van den Brakel, *Statistics Netherlands*  
 J.M. Brick, *Westat Inc*  
 P.D. Brick, *Westat Inc*  
 C. Calder, *Ohio State University*  
 P. Cantwell, *U.S. Bureau of the Census*  
 R. Chambers, *Centre for Statistical and Survey Methodology*  
 P. Dick, *Statistique Canada*  
 A.H. Dorfman, *U.S. Bureau of Labour Statistics*  
 G. Dubreuil, *Statistique Canada*  
 M. Elliott, *University of Michigan*  
 J.L. Eltinge, *U.S. Bureau of Labor Statistics*  
 G. Falk, *University of Virginia*  
 M. Fay, *National Institute of Allergy and Infectious Diseases*  
 W.A. Fuller, *Iowa State University*  
 J. Gambino, *Statistique Canada*  
 N. Ganesh, *National Opinion Research Center at the University of Chicago*  
 S. Ghosh, *Alberta Health Services-Cancer care*  
 C. Girard, *Statistique Canada*  
 S. Godbout, *Statistique Canada*  
 R. Griffin, *U.S. Census Bureau*  
 D. Haziza, *Université de Montréal*  
 Y. He, *Harvard Medical School*  
 S. Heeringa, *University of Michigan*  
 M. Hidiroglou, *Statistique Canada*  
 S. Holan, *University of Missouri*  
 A. Holmberg, *Statistics Sweden*  
 D. Hubble, *Westat Inc*  
 B. Hülliger, *University of Applied Sciences Northwestern Switzerland*  
 J. Jones, *Office for National Statistics, United Kingdom*  
 D. Judkins, *Westat Inc*  
 D. Kasprzyk, *Mathematica Policy Research*  
 P. Kelly, *Statistique Canada*  
 P. Kott, *RTI International*  
 P. Lahiri, *JPSM, University of Maryland*  
 M.D. Larsen, *George Washington University*  
 P. Lavallée, *Statistique Canada*  
 J. Legg, *Amgen Inc, USA*  
 R. Little, *University of Michigan*  
 A. Zaslavsky, *Harvard Medical School*  
 P.A. Zandbergen, *University of New Mexico*  
 W. Yung, *Statistique Canada*  
 C. Wu, *University of Waterloo*  
 K.M. Wolter, *Iowa State University*  
 W.E. Winkler, *U.S. Census Bureau*  
 V.J. Verma, *Università degli Studi di Siena*  
 C. Tucker, *U.S. Bureau of Labor Statistics*  
 D. Toth, *U.S. Bureau of Labor Statistics*  
 S. Thompson, *Simon Fraser University*  
 M. Thompson, *University of Waterloo*  
 A. Thèberge, *Statistique Canada*  
 L. Stokes, *Southern Methodist University*  
 D. Steel, *University of Wollongong*  
 E. Stasny, *Ohio State University*  
 T.M.F. Smith, *University of Southampton, UK*  
 P. Smith, *Office for National Statistics*  
 P. do N. Silva, *Escola Nacional de Ciências Estatísticas*  
 F.J. Scheuren, *National Opinion Research Center*  
 N. Schenker, *National Center for Health Statistics*  
 C.-E. Sæmndal, *Université de Montréal*  
 N. Salvati, *Università di Pisa*  
 J. Rytén  
 S. Rubin-Bleuer, *Statistique Canada*  
 L.-P. Rivest, *Université Laval*  
 J. Reiter, *Duke University*  
 J.N.K. Rao, *Carleton University*  
 M. Pratesi, *Università di Pisa*  
 N.G.N. Prasad, *University of Alberta*  
 D. Pfeffermann, *Hebrew University*  
 Z. Patak, *Statistique Canada*  
 J. Opsomer, *Colorado State University*  
 A. Nigam, *Institute of Applied Statistics and Development Studies*  
 S.F. Nielsen, *Copenhagen Business School, Denmark*  
 G. Nathan, *Hebrew University*  
 T. Mulcahy, *National Opinion Research Center*  
 K. Miller, *U.S. National Center for Health Statistics*  
 E. Martin, *U.S. Census Bureau*  
 J. Maples, *U.S. Census Bureau*  
 D.J. Malec, *U.S. Census Bureau*  
 M. Maia, *Catholic University of Portugal*  
 P. Lynn, *University of Essex*  
 S. Lohr, *Arizona State University*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2010 : Céline Ethier de la Division de la recherche et de l'innovation en statistique, Christine Cousineau et Teresa Jewell de la Division des méthodes d'enquêtes auprès des ménages, Nick Budko et Sophie Chartier de la Division des méthodes d'enquêtes auprès des entreprises, Matthew Belyea, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Darguise Pellerin de la Division de la diffusion, et Jeff Jodoin de la Division des services à la clientèle.





Entropie des plans (échantillonnage de Bernoulli ; échantillonnage aléatoire simple)					
Tableau 1					
N	$\pi = 0,1$	$\pi = 0,2$	$\pi = 0,3$	$\pi = 0,4$	$\pi = 0,5$
10	(3,3 ; 2,3)	(5 ; 3,8)	(6,1 ; 4,8)	(6,7 ; 5,3)	(6,9 ; 5,5)
100	(32,5 ; 30,5)	(50 ; 47,7)	(61,1 ; 58,6)	(67,3 ; 64,8)	(69,3 ; 66,8)
1 000	(325,1 ; 321,9)	(500,4 ; 496,9)	(610,9 ; 607,3)	(673, ; 669,4)	(693,1 ; 689,5)
10 000	(3 250,8 ; 3 246,5)	(5 004 ; 4 999,4)	(6 108,6 ; 6 103,9)	(6 730,1 ; 6 725,3)	(6 931,5 ; 6 926,6)
100 000	(32 508,3 ; 32 502,8)	(50 040,2 ; 50 034,5)	(61 086,4 ; 61 080,5)	(67 301,2 ; 67 295,2)	(69 314,7 ; 69 308,7)
1 000 000	(325 083 ; 325 076)	(500 402 ; 500 396)	(610 864 ; 610 857)	(673 012 ; 673 005)	(693 147 ; 693 140)

Nous remercions un rédacteur associé et un examinateur de leurs commentaires constructifs. Les travaux de David Haziza ont été financés en partie par des bourses de recherche du Conseil de recherches en sciences naturelles et en génie du Canada.

### Bibliographie

Abramowitz, M., et Stegun, I.A. (1964). *Handbook of Mathematical Functions*. New York : Dover.

Brewer, K.R.W., et Donadio, M.E. (2003). La variance sous grande entropie de l'estimateur de Horvitz-Thompson. *Techniques d'enquête*, 29, 213-220.

Chen, S.X., Dempster, A.P. et Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Rapport technique, CREST-ENSAI, Rennes.

Haziza, D., Mecatti, F. et Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron*, 66, 91-108.

Henderson, T. (2006). Estimating the variance of the Horvitz-Thompson estimator. Thèse de maîtrise, School of Finance and Applied Statistics, The Australian National University.

Matei, A., et Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21, 4, 543-570.

Tillé, Y. (2006). *Sampling Algorithms*. New York : Springer.

est celui connu sous le nom d'échantillonnage de Poisson conditionnel (EPC) ; (voir Chen, Dempster et Liu 1994 ; Deville 2000 ; Tillé 2006). Le plan EPC peut être mis en œuvre en sélectionnant des échantillons à plusieurs reprises conformément à l'échantillonnage de Poisson jusqu'à l'obtention de la taille d'échantillon souhaitée, disons  $n$ . Quand  $\pi_k = n/N$  pour tout  $k \in U$ , le plan EPC se réduit à l'échantillonnage aléatoire simple sans remise :

$$p_{\text{cas}}(s) = \frac{n}{N} \binom{n}{n}^{-1}$$

pour tout  $s \in \tilde{Q}$ . Il découle de (1) que l'entropie de l'échantillonnage aléatoire simple est donnée par

$$(4) \quad I(p_{\text{cas}}) = \log N! - \log n! - \log(N - n)!.$$

Autrement dit, l'échantillonnage aléatoire simple sans remise est le plan à entropie maximale dans la classe des plans d'échantillonnage équiprobable avec taille fixe.

Les plans d'échantillonnage  $n$ 'ont pas tous une entropie

élevée. Par exemple, le plan d'échantillonnage systématique

1 sur  $G$  possède une très faible entropie. Ici, on suppose que le nombre d'échantillons,  $G = N/n$ , est un entier.

Puisque  $p_{\text{sys}}(s) = 1/G$  pour tout  $s \in \tilde{Q}$ , l'entropie de

l'échantillonnage systématique est donnée par

$$I(p_{\text{sys}}) = \log N - \log n,$$

qui est beaucoup plus faible que (4), surtout pour les

grandes valeurs de  $N$ .

## 2. Résultat principal

À la présente section, nous comparons l'entropie de l'échantillonnage de Bernoulli à celle de l'échantillonnage aléatoire simple sans remise. Puisque le support des plans d'échantillonnage de Bernoulli est beaucoup plus grand que celui des plans d'échantillonnage aléatoire simple sans remise, nous nous attendons à ce que l'entropie de l'échantillonnage de Bernoulli soit beaucoup plus élevée que celle de l'échantillonnage aléatoire simple sans remise. Le tableau 1 donne l'entropie de l'échantillonnage aléatoire simple et de l'échantillonnage de Bernoulli pour diverses valeurs de  $N$  et de  $\pi$ . Étonnamment, nous constatons que l'entropie des deux plans d'échantillonnage pour les mêmes probabilités d'inclusion et la même taille d'échantillon est à peu près la même. L'examen du tableau 1 montre clairement que les deux plans d'échantillonnage ont des entropies similaires, même pour des tailles de population modestes (par exemple,  $N = 100$ ), indépendamment de la valeur de  $\pi$ . Ce résultat est un peu curieux, étant donné la réduction

$$\lim_{N \rightarrow \infty} \frac{I(p_{\text{cas}})}{I(p_{\text{bern}})} = 1.$$

aléatoire simple sans remise. Résultat 1. Soit  $I(p_{\text{bern}})$  et  $I(p_{\text{cas}})$  les entropies de l'échantillonnage de Bernoulli et de l'échantillonnage aléatoire simple sans remise, respectivement, données par (3) et (4). Alors,

importante du nombre d'échantillons possibles lorsque l'on fixe la taille de l'échantillon. En effet, rappelons que la taille du support est  $\binom{n}{N}$  pour l'échantillonnage aléatoire simple sans remise, tandis qu'elle est égale à  $2^N$  pour l'échantillonnage de Bernoulli. Par exemple, pour  $N = 100$  et  $n = 20$ , la taille du support pour l'échantillonnage aléatoire simple sans remise est égale à  $\binom{100}{20} \approx 5,36 \times 10^{20}$ , tandis qu'elle est égale à  $2^{100} \approx 1,26 \times 10^{30}$  pour l'échantillonnage de Bernoulli. Autrement dit, la taille du support de l'échantillonnage de Bernoulli est environ  $2,36 \times 10^9$  fois plus grande que celle du support de l'échantillonnage aléatoire simple sans remise.

*Preuve.* En partant de la formule de Stirling (voir Abramowitz et Stegun 1964, page 257)

$$\lim_{n \rightarrow \infty} \frac{n \log n - n}{\log n!} = 1,$$

nous obtenons

$$\lim_{\substack{N \rightarrow \infty \\ n \rightarrow \infty \\ N-n \rightarrow \infty}} \frac{N \log N - n \log n - (N - n) \log(N - n)}{N \log \binom{n}{N}} = 1,$$

d'où nous obtenons

$$\lim_{N \rightarrow \infty} \frac{\log \binom{n}{N} - N \log \pi}{-N(1 - \pi) \log(1 - \pi) - N \pi \log \pi} = 1.$$

## 3. Conclusion

Dans la présente note, nous avons montré que l'échantillonnage de Bernoulli et l'échantillonnage aléatoire simple sans remise ont des entropies fort semblables, même pour des tailles de population modestes. Nous conjecturons que nous ferions la même constatation en comparant le plan d'échantillonnage de Poisson et le plan d'échantillonnage de Poisson conditionnel pour un ensemble donné de probabilités d'inclusion de premier ordre. Toutefois, la preuve de ce résultat semble être considérablement plus complexe.



# Une propriété intéressante de l'entropie de certains plans d'échantillonnage

## Yves Tillé et David Haziza<sup>1</sup>

### Résumé

Dans cette note brève, nous montrons que l'échantillonnage aléatoire sans remise et l'échantillonnage de Bernoulli ont à peu près la même entropie quand la taille de la population est grande. Nous donnons un exemple empirique en guise d'illustration.

Mots clés : Échantillonnage de Poisson conditionnel ; entropie ; échantillonnage aléatoire simple ; échantillonnage de Poisson.

### 1. Introduction

Considérons une population finie de taille  $N$  et soit  $U = \{1, \dots, k, \dots, N\}$  l'ensemble des étiquettes de cette population. Un échantillon  $s$  est un sous-ensemble de  $U$  et un plan d'échantillonnage est une loi de probabilité  $p(\cdot)$  sur les sous-ensembles de  $U$  telle que  $p(s) \geq 0$  pour tout  $s \subset U$ , et

$$\sum_{s \subset U} p(s) = 1.$$

Soit  $\pi_k = P(k \in s)$  la probabilité d'inclusion de premier ordre de l'unité  $k$  dans l'échantillon :

$$\pi_k = \sum_{s \subset U, k \in s} p(s).$$

De même, soit  $\pi_{k\ell} = P(k \in s \text{ et } \ell \in s)$  la probabilité d'inclusion de deuxième ordre des unités  $k$  et  $\ell$  dans l'échantillon :

$$\pi_{k\ell} = \sum_{s \subset U, k, \ell \in s} p(s).$$

L'entropie d'un plan d'échantillonnage  $p(\cdot)$ , désignée par  $I(p)$ , est définie comme étant

$$I(p) = - \sum_{s \in \tilde{Q}} p(s) \log p(s), \quad (1)$$

où  $\tilde{Q} = \{s | p(s) > 0\}$  est le support du plan d'échantillonnage  $p(\cdot)$ . Un plan d'échantillonnage possède une entropie élevée quand le degré d'incertitude ou de surprise est grand en ce qui concerne l'échantillon qui sera sélectionné. Autrement dit, quand un plan d'échantillonnage possède une entropie élevée, il est très difficile de prédire le type d'échantillon que l'on obtiendra. De nombreux plans

d'échantillonnage utilisés en pratique sont des plans à entropie élevée. L'échantillonnage systématique, dont l'entropie est très faible, est une exception notable. Le concept d'entropie est utile dans le contexte de l'estimation de la variance. Lorsque l'entropie d'un plan d'échantillonnage est élevée, il est possible d'obtenir une approximation des probabilités d'inclusion de deuxième ordre,  $\pi_{k\ell}$ , en fonction des probabilités d'inclusion de premier ordre, ce qui simplifie considérablement le problème d'estimation de la variance dans le contexte de l'échantillonnage avec probabilités inégales ; voir, par exemple, Brewer et Donadio (2003), Matei et Tillé (2005), Henderson (2006) et Haziza, Mecatti et Rao (2008).

Il est bien connu que le plan d'échantillonnage à entropie maximale est l'échantillonnage de Poisson :

$$p^{\text{poiss}}(s) = \prod_{k \in U, s} \pi_k \prod_{k \notin U, s} (1 - \pi_k) \quad (2)$$

pour tout  $s \in \tilde{Q}$  ; voir par exemple, Tillé (2006). Un cas particulier de l'échantillonnage de Poisson est l'échantillonnage de Bernoulli, que l'on obtient à partir de (2) en posant que  $\pi_k = \pi \in (0, 1)$ , ce qui mène à

$$p^{\text{bern}}(s) = \pi^{n_s} (1 - \pi)^{N - n_s}, \text{ pour tout } s \subset U,$$

où  $n_s$  est la taille aléatoire de  $s$ . En utilisant (1) et en notant que  $\sum_{s \in \tilde{Q}} n_s p(s) = N\pi$ , l'entropie de l'échantillonnage de Bernoulli est donnée par

$$I(p^{\text{bern}}) = -N(1 - \pi) \log(1 - \pi) - N\pi \log \pi. \quad (3)$$

qui est maximale quand  $\pi = 1/2$ . Dans ce cas, nous avons  $I(p^{\text{bern}}) = N \log 2$ . Si nous limitons à la classe des plans d'échantillonnage avec taille fixe et probabilités d'inclusion de premier ordre  $\pi_k, k \in U$ , le plan dont l'entropie est maximale

<sup>1</sup> Yves Tillé, Institut de Statistique, Université de Neuchâtel, Neuchâtel, QC, Canada, H3C 3J7. Courriel : haziza@dms.umontreal.ca.  
David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, QC, Canada, H3C 3J7.

- O'Neil, M.J. (1979). Estimating the nonresponse bias due to refusals in the telephone surveys. *Public Opinion Quarterly*, 43, 218-232.
- SAS Institute Inc. (2004). SAS version 9.1. Cary, NC: SAS Institute, USA.
- Troiano, R.P., Bertigan, D., Dodd, K.W., Masse, L.C., Tillet, T. et McDowell, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise*, 40, 181-8.
- Voigt, L.F., Koepsell, T.D. et Daling, J.R. (2003). Characteristics of Telephone Survey Respondents According to Willingness to Participate. *American Journal of Epidemiology*, 157, 66-73.
- Qayad, M.G., Balluz, L. et Garvin, W. (2009). Does continuing data collection beyond one month improve the completion and response rates in behavioral risk factor surveillance system survey? *Survey Practice Feb 2009*. Adresse URL <http://surveypractice.org/2009/02/>.
- Rao, R.S., Link, M.W., Battaglia, M.P., Frankel, M.R., Giambro, P. et Mokdad, A.H. (2005). Assessing representativeness in RDD surveys: coverage and non-response in the Behavioral Risk Factor Surveillance System. Minneapolis: Joint Statistical Meetings. Adresse URL <http://www.amstat.org/sections/SRMS/Proceedings/y2005/Files/JSMSM2005-000190.pdf>.

#### 4. Discussion

Notre étude a révélé des différences significatives entre les répondants hâtifs et les répondants tardifs en ce qui concerne les caractéristiques démographiques ainsi que certains comportements posant un risque pour la santé et maladies ou problèmes de santé chroniques. Cela montre que la composition des deux groupes de répondants diffère en ce qui a trait à ces attributs. Les différences observées pourraient être attribuables à la difficulté de joindre des personnes qui travaillent de longues heures et à leur absence du domicile.

Chez les répondants tardifs, la plus forte probabilité d'avoir un revenu élevé, d'être hispanique, d'être jeune (entre 18 et 44 ans), de bénéficier d'une couverture des soins de santé, d'avoir un niveau de scolarité inférieur au diplôme d'études secondaires et de déclarer un bon état de santé général correspond aux caractéristiques décrites des personnes qui travaillent et des travailleurs en santé (Li et Sung 1999), (O'Neil 1979). Leur probabilité significative ment plus faible de déclarer souffrir d'hypertension ou de diabète et d'être obèse vient étayer cette description. Toutefois, certains comportements posant un risque affichent un profil différent chez les répondants tardifs. Ces derniers sont moins susceptibles de respecter les lignes directrices recommandées en matière d'activité physique modérée ou vigoureuse et de consommation quotidienne de fruits et de légumes, ce qui peut être relié au fait que les répondants tardifs ont de longues heures de travail et peu d'accès à des aliments sains.

Les personnes à revenu élevé, dont la plupart sont des Blancs non hispaniques et les personnes à faible revenu, dont la plupart sont des Hispaniques et des Noirs non hispaniques, peuvent passer de longues heures dans leur milieu de travail et être moins susceptibles d'être à la maison pour prendre un appel d'enquête (Voigt, Koepsell et Daling 2003). En outre, les données du BRFSS montrent que les intervieweurs font plus d'appels aux répondants tardifs, en moyenne presque trois fois plus qu'aux répondants hâtifs, ce qui témoigne de la difficulté de les joindre durant la période d'enquête de 31 jours. Les raisons des longues heures de travail peuvent être différentes pour les deux groupes de revenu. Les Hispaniques, les Noirs non hispaniques et les jeunes peuvent occuper des emplois faiblement rémunérés et donc être obligés de travailler de longues heures pour gagner leur vie, tandis que les personnes à revenu élevé peuvent occuper des emplois qui les obligent de rester au travail après les heures normales de travail.

Il faut examiner de près les estimations épidémiologiques et décollant de la surveillance fondées seulement sur les répondants hâtifs ou tardifs pour déceler tout biais éventuel avant de faire des généralisations. Le pourcentage d'interviews menées après 31 jours est actuellement petit (5 %) et leur exclusion de l'analyse peut n'avoir aucun effet sur les estimations aux niveaux national et de l'État. Toutefois, étant donné que les proportions de répondants tardifs

#### Bibliographie

Les auteurs voudraient remercier les coordonnateurs du BRFSS ainsi que les entrepreneurs chargés de la cueillette des données au niveau des états.

#### Remerciements

Malgré ces limites, l'étude montre que les répondants tardifs diffèrent significativement à de nombreux égards des répondants hâtifs. Étant donné que la proportion de répondants tardifs peut augmenter à l'avenir, il faut examiner soigneusement l'incidence des répondants tardifs sur les estimations décollant de la surveillance.

États.

Malgré ces limites, l'étude montre que les répondants tardifs diffèrent significativement à de nombreux égards des répondants hâtifs. Étant donné que la proportion de répondants tardifs peut augmenter à l'avenir, il faut examiner soigneusement l'incidence des répondants tardifs sur les estimations décollant de la surveillance.

Notre étude présente quelques limites. Le BRFSS utilise la méthode de composition aléatoire pour sélectionner les numéros de téléphone, d'où un éventuel biais de couverture (Rao, Link, Battaglia, Frankel, Giambo et Mokdad 2005; Frankel, Link, Battaglia, Frankel, Hoaglin, Smith, Wright et Khare 2003). Les renseignements recueillis sont auto-déclarés et pourraient être sujets à un biais de rappel dans certaines estimations des maladies et des comportements posant un risque (Troiano, Berrigan, Dodd, Masse, Tillet et McDowell 2008; CDC 2004). En outre, nous avons exclu deux États de notre analyse (le Michigan et la Louisiane), et il faut procéder avec prudence en extrapolant les résultats à ces États.

Centers for Disease Control and Prevention (2009a). BRFSS Turning Information into Public Health, adresse URL <http://www.cdc.gov/btrfs/about.htm>.

Centers for Disease Control and Prevention (2004). Access to health-care and preventive services among Hispanics and non-Hispanics – United States, 2001-2002. *Morbidity and Mortality Weekly Report*, 53, 937-941.

Diehr, P., Cain, K., Connell, F. et Vollim, E. (1990). What is too much variation? The null hypothesis in small-area analysis. *Health Services Research*, 24, 741-771.

Frankel, M.R., Srinath, K.P., Hoaglin, D.C., Battaglia, M.P., Smith, P.J., Wright, R.A. et Khare, M. (2003). Adjustments for non-telephone bias in random-digit-dialing surveys. *Statistics in Medicine*, 22, 1611-1626.

Li, C.Y., et Sung, F.C. (1999). A review of the healthy worker effect in occupational epidemiology. *Occupational Medicine*, 49, 225-229.



Tableau 2

Différences en pourcentage entre les répondants hâtifs et les répondants tardifs selon les comportements posant un risque pour la santé,

BRFSS 2007

Facteurs de risque				
Durée de la période de collecte				
Répondants hâtifs*	Répondants tardifs**	Différence	(Hâtifs-tardifs)	Valeur p
(N = 394 427)	(N = 22 247)	%	%	
Consommation excessive d'alcool	11,1	11,8	-0,7	0,261
Oui	86,9	82,8	4,1	
Non	1,9	5,4	-3,4	
Consommation de cigarettes	18,3	17,5	0,9	0,182
Fumeurs	81,3	82,1	-0,8	
Non-fumeurs	0,4	0,5	0,0	
Respectent les recommandations en matière d'activité physique	43,4	41,8	1,7	0,000
modérée ou vigoureuse	35,4	31,8	3,6	
Activité physique insuffisante	14,3	11,3	3,0	
Pas d'activités physiques	6,9	15,2	-8,3	
Consommation de fruits et de légumes	25,0	21,9	3,1	0,000
Consommation ≥ 5 fois par jour	73,0	69,7	3,3	
Consommation < 5 fois par jour	2,0	8,5	-6,4	
Inconnu				

\*Ont participé à l'enquête dans les 31 jours.

\*\*Ont participé à l'enquête après 31 jours.

Tableau 3

Différence en pourcentage entre les répondants hâtifs et les répondants tardifs selon les maladies et problèmes de santé chroniques,

BRFSS 2007

Maladies/problèmes de santé chroniques				
Durée de la période de collecte				
Répondants hâtifs*	Répondants tardifs**	Différence	(Hâtifs-tardifs)	Valeur p
(N = 394 427)	(N = 22 247)	%	%	

Maladies cérébrovasculaires et maladies cardiovasculaires

Oui	5,9	4,9	1,0	0,177
Non	93,6	94,7	-1,1	
Inconnu	0,5	0,4	0,1	
Angine de poitrine	6,0	4,5	1,5	0,053
Oui	93,1	94,7	-1,6	
Non	0,9	0,8	0,1	
Inconnu	3,8	2,8	1,0	0,183
Non	95,9	97,0	-1,1	
Inconnu	0,3	0,2	0,1	
Autres maladies/problèmes de santé :				
Taux élevé de cholestérol	57,0	60,8	-3,8	0,000
Oui	42,3	38,4	3,8	
Non	57,0	60,8	-3,8	
Inconnu	0,8	0,8	0,0	
Hypertension	35,8	30,1	5,8	0,000
Oui	64,0	69,8	-5,8	
Non	0,2	0,2	0,0	
Inconnu	11,2	9,4	1,8	0,010
Oui	0,9	1,2	-0,2	
Oui-Crossesse	86,4	88,2	-1,9	
Non	1,4	1,2	0,2	
Inconnu	0,1	0,1	0,0	
Ashme	8,7	7,7	1,0	0,158
Actuellement	3,8	4,0	-0,2	
Dans le passé	86,9	87,8	-0,8	
Jamais	0,6	0,6	0,1	
Inconnu	34,5	35,5	-1,1	
Poids normal	26,0	23,6	2,4	0,000
Embonpoint	4,5	6,2	-1,7	
Obésité	34,5	35,5	-1,1	
Inconnu	26,0	23,6	2,4	

\*Ont participé à l'enquête dans les 31 jours.

\*\*Ont participé à l'enquête après 31 jours.

de déclarer leur race ou origine ethnique comme étant hispanique, d'avoir un revenu annuel supérieur ou égal à 50 000 \$, d'avoir moins de 45 ans, d'avoir un niveau de scolarité inférieur au diplôme d'études secondaires, d'avoir accès à une couverture des soins de santé et à déclarer être en bonne santé. La valeur absolue des différences significatives dans les variables ci-dessus se situe entre 1,3 % et 7,6 %. Le pourcentage de « Inconnu » dans la variable de couverture des soins de santé est de 21 % pour les répondants tardifs et de 30 % pour les répondants hâtifs. La différence entre les répondants hâtifs et les répondants tardifs demeure significative même sous l'hypothèse que le pourcentage de ceux ayant accès à la couverture des soins de santé chez ceux dont la situation est inconnue est comparable à celui dont la situation est connue.

Une différence significative entre les répondants hâtifs et les répondants tardifs s'observe également en ce qui concerne les comportements posant un risque pour la santé (tableau 2). Comparativement aux répondants hâtifs, les répondants tardifs étaient significativement moins susceptibles

**Tableau 1**  
Différences en pourcentage entre les répondants hâtifs et les répondants tardifs selon les caractéristiques démographiques, la couverture des soins de santé et l'état de santé général, BRFSS 2007

Caractéristiques démographiques				Durée de la période de collecte	
		Répondants hâtifs* (N = 394 427)	Répondants tardifs** (N = 22 247)	Différence (Hâtifs-tardifs)	Valeur p
Sexe		%	%	%	
Femmes		62,8	60,2	2,5	0,000
Hommes		37,3	39,8	-2,5	
Race					
Blanche, non hispanique		79,1	71,5	7,6	0,000
Noire, non hispanique		7,3	8,2	-0,9	0,168
Hispanique		7,1	13,5	-6,4	0,000
Autres		5,5	5,8	-0,3	0,635
Inconnu		1,0	1,0	0,0	0,977
Revenu					
<15 000		9,7	8,7	1,0	0,146
15 à 34 999		26,1	24,3	1,8	0,004
35 à 49 999		14,1	13,4	0,8	0,252
50 000+		36,6	39,7	-3,1	0,000
Inconnu		13,5	14,0	-0,4	0,496
Âge					
18 à 24		3,6	4,9	-1,3	0,025
25 à 44		25,7	33,3	-7,6	0,000
45 à 64		40,9	40,6	0,3	0,612
65+		29,0	20,2	8,8	0,000
Inconnu		0,8	1,0	-0,1	0,827
Niveau de scolarité					
Niveau inférieur au diplôme d'études secondaires		10,3	12,3	-2,0	0,001
Diplôme d'études secondaires		30,6	28,7	1,9	0,001
Niveau supérieur aux études secondaires		58,8	58,2	0,6	0,177
Inconnu		0,3	0,8	-0,5	0,264
Couverture des soins de santé (<65 ans)					
Oui		59,3	65,4	-6,2	0,000
Non		10,8	13,2	-2,5	
Inconnu		30,0	21,4	8,6	
État de santé					
Bonne santé		80,1	81,8	-1,7	0,000
Santé passable ou mauvaise		19,4	17,6	1,8	
Inconnu		0,5	0,6	-0,1	

\*Ont participé à l'enquête dans les 31 jours.  
\*\*Ont participé à l'enquête après 31 jours.

surveillance à l'avenir s'impose. La présente étude vise à déterminer si les répondants qui ont participé aux interviews dans les 31 jours et ceux qui y ont participé après 31 jours diffèrent pour ce qui est des caractéristiques démographiques, des comportements posant un risque et des maladies et problèmes de santé chroniques.

## 2. Méthodes

Nous avons utilisé les données du BRFSS 2007, qui est une enquête téléphonique par composition aléatoire menée auprès de la population civile ne vivant pas en établissement aux États-Unis. Nous avons divisé la durée de l'interview en deux périodes, soit de 0 à 31 jours et de plus de 31 jours. Les répondants ont été répartis en deux groupes selon qu'ils avaient participé à l'interview dans les 31 jours (répondants hâtifs) ou qu'ils y avaient participé après 31 jours (répondants tardifs).

Les caractéristiques démographiques incluses étaient le sexe, la race, le revenu et l'âge. La race comprenait quatre groupes : blanche, non hispanique ; noire, non hispanique ; hispanique ; et autre. Les niveaux de scolarité étaient au nombre de trois : pas de diplôme d'études secondaires, diplôme d'études secondaires et niveau supérieur aux études secondaires. Les catégories de revenu étaient les suivantes : moins de 15 000 \$, de 15 000 \$ à 34 999 \$, de 35 000 \$ à 49 999 \$ et de 50 000 \$ ou plus. Les catégories d'âge étaient 18 à 24 ans, 25 à 44 ans, 45 à 64 ans et 65 ans et plus. Les répondants de moins de 65 ans non couverts par un régime de soins de santé (y compris un régime d'assurance-santé, des régimes payés comme ceux de l'OSSI ou un régime gouvernemental comme le régime d'assurance-maladie Medicare) ont été considérés comme n'étant pas couverts par un régime de soins de santé. L'état de santé général a été dichotomisé en bonne santé (excellent, très bonne ou bonne) et santé passable ou mauvaise.

Les comportements posant un risque pour la santé examinés étaient la consommation abusive d'alcool, l'usage du tabac au moment de l'enquête, (l'insuffisance de) l'activité physique et la consommation (insuffisante) de fruits et de légumes. La consommation abusive d'alcool a été définie comme le fait d'avoir consommé au moins cinq verres d'alcool pour les hommes et au moins quatre verres d'alcool pour les femmes à au moins une occasion au cours du mois qui a précédé l'enquête. Les répondants qui avaient fumé 100 cigarettes ou plus au cours de leur vie et qui fumaient quotidiennement ou certains jours ont été classés comme des fumeurs. Pour l'activité physique, les catégories étaient les suivantes : respectent les recommandations en matière d'activité physique, activité physique insuffisante et ne participent pas à des activités physiques. Les répondants qui consommaient au moins cinq portions de fruits et de légumes chaque jour ont été classés comme respectant les recommandations en matière de consommation de fruits et de légumes.

Les maladies ou problèmes de santé chroniques inclus étaient les maladies cardiovasculaires et cardiaques, l'hypertension, le niveau de cholestérol élevé, le diabète, l'asthme et l'embonpoint ou l'obésité. Les répondants ont été considérés comme ayant un infarctus du myocarde, souffrant d'angine de poitrine, ayant eu un accident vasculaire cérébral ou faisant de l'hypertension artérielle si un médecin, une infirmière ou un autre professionnel de la santé leur avait déjà dit qu'ils avaient l'un de ces problèmes. Les répondants ont été classés comme souffrant d'hypertension si leur cholestérol avait été vérifié et un professionnel de la santé leur avait dit que leur taux de cholestérol dans le sang était élevé. Les répondants ont été classés comme souffrant de diabète si un médecin leur avait déjà dit qu'ils étaient atteints de diabète. L'asthme était auto-déclaré et diagnostiqué par un médecin ou un professionnel de la santé ; trois catégories étaient établies à cet égard, selon que le répondant faisait de l'asthme actuellement, avait fait de l'asthme dans le passé ou n'avait jamais fait d'asthme. La taille et le poids auto-déclarés ont été utilisés pour calculer l'indice de masse corporelle (IMC) ( $IMC = \text{poids} [kg] / \text{taille} [m]^2$ ). Les participants ont été classés comme faisant de l'embonpoint si leur IMC était supérieur ou égal à 25 kg/m<sup>2</sup> et comme étant obèses si leur IMC était supérieur ou égal à 30 kg/m<sup>2</sup>.

Nous avons estimé les différences en pourcentage entre les répondants hâtifs et les répondants tardifs selon les caractéristiques démographiques, les comportements influant sur la santé et les maladies ou problèmes de santé chroniques. Pour exécuter l'analyse, nous sommes servis de SUDAAN et de SAS (SAS Institute Inc., Cary, Caroline du Nord, E.-U., 2004).

## 3. Résultats

Dans le cadre de l'enquête BRFSS 2007, 430 912 interviews ont été menées aux États-Unis. Nous avons exclu 14 189 enregistrements de deux États (le Michigan et la Louisiane) et 49 cas dans lesquels de l'information était manquante. Nous avons analysé les 416 674 répondants restants, dont 394 427 (95 %) étaient des répondants hâtifs et 22 247 (5 %), des répondants tardifs. Nous avons estimé les différences en pourcentage pondérées et non pondérées entre les répondants hâtifs et les répondants tardifs. Les différences absolues entre les pourcentages pondérés et non pondérés dans les variables examinées se situent entre 0,6 % et 2,6 %, sauf pour les Blancs non hispaniques pour lesquels la différence absolue est de 7 %. Aux fins de la présente étude, nous présentons les résultats de l'analyse non pondérée.

Des différences significatives s'observent entre les répondants hâtifs et les répondants tardifs en ce qui a trait aux variables démographiques, d'accès à la couverture des soins de santé et d'état de santé général (tableau 1). Comparativement aux répondants hâtifs, les répondants tardifs étaient significativement plus susceptibles d'être de sexe masculin,



# Différences entre les répondants et durée de la période de collecte des données dans le Behavioral Risk Factor Surveillance System

## Mohamed G. Qayad, Pranesh Chowdhury, Shaohua Hu et Lina Balluz

### Résumé

Le ralentissement économique aux États-Unis pourrait rendre incertain le maintien de stratégies coûteuses dans les opérations des enquêtes. Dans le Behavioral Risk Factor Surveillance System (BRFSS), une période de collecte de données mensuelle de 31 jours seulement pourrait être une solution de rechange moins coûteuse. Toutefois, elle pourrait exclure une partie des interviews menées après 31 jours (répondants tardifs) et les caractéristiques de ces répondants pourraient être différentes à de nombreux égards de celles des répondants qui ont participé à l'enquête dans les 31 jours (répondants hâtifs). Nous avons laché de déterminer s'il existe entre les répondants hâtifs et les répondants tardifs des différences d'ordre démographique ou en ce qui a trait à la couverture des soins de santé, à l'état de santé général, aux comportements posant un risque pour la santé et aux maladies ou problèmes de santé chroniques. Nous avons utilisé les données du BRFSS 2007, où un échantillon représentatif de la population adulte aux États-Unis ne vivant pas en établissement a été sélectionné au moyen d'une méthode de composition aléatoire. Les répondants tardifs étaient significativement plus susceptibles d'être de sexe masculin ; de déclarer leur race ou origine ethnique comme étant hispanique ; d'avoir un revenu annuel de plus de 50 000 \$ ; d'avoir moins de 45 ans ; d'avoir un niveau de scolarité inférieur au diplôme d'études secondaires ; de bénéficier d'une couverture des soins de santé ; d'être significativement plus susceptibles de déclarer être en bonne santé ; d'être observées entre les répondants hâtifs et les répondants tardifs dans les estimations d'enquête pourraient influer à peine sur les estimations nationales et au niveau de l'État. Étant donné que la proportion de répondants tardifs pourrait augmenter à l'avenir, il y a lieu d'examiner son incidence sur les estimations découlant de la surveillance avant de l'exclure de l'analyse. Dans l'analyse portant sur les répondants tardifs, il devrait suffire de combiner plusieurs années de données pour produire des estimations fiables.

Mois clés : BRFSS ; répondants ; différences ; durée de la période de collecte des données.

### 1. Introduction

Le Behavioral Risk Factor Surveillance System (BRFSS) est une enquête par téléphone menée auprès des ménages dans les divers États des États-Unis et dans ses territoires qui vise à surveiller les comportements posant un risque pour la santé et les maladies et problèmes de santé chroniques chez la population adulte ne vivant pas en établissement (Centers for Disease Control and Prevention [CDC] 2009a, BRFSS Turning Information into Public Health, <http://www.cdc.gov/brfss/about.htm>). Il s'agit de la plus vaste enquête par téléphone au monde et elle est menée par les 50 États, le district fédéral de Columbia et les territoires des États-Unis, de concert avec les CDC. L'enquête est menée de façon continue tout au long de l'année.

Les CDC attribuent les échantillons (numéros de téléphone) aux États sur une base trimestrielle. Au niveau de l'État, les échantillons sont divisés en 12 listes mensuelles aux fins opérationnelles. Des intervieweurs ayant reçu une formation pertinente placent un appel à chaque numéro de téléphone échantillonné. Un code de décision est attribué après chaque appel. Les États et leurs entrepreneurs sont tenus de prendre des décisions finales relativement à leurs échantillons mensuels sur lesquels des données sont diffusées au cours du même mois. Plus de 90 % des échantillons

mensuels et des interviews menées font l'objet d'une décision finale dans les 31 jours. Les États continuent de travailler sur leurs échantillons restants par la suite (Qayad, Balluz et Garvin 2009).

Les ralentissements économiques pourraient entraîner des compressions budgétaires pouvant avoir un effet négatif sur les opérations des enquêtes des États et des organismes d'enquête. Par ailleurs, des circonstances imprévues justifient la recherche de stratégies opérationnelles de rechange. Mettre fin à la collecte des données à la fin de chaque mois pourrait constituer une solution de rechange rentable. Toutefois, une période de collecte des données ne dépassant pas un mois exclut les interviews menées après 31 jours. Ces exclusions pourraient influer sur la variabilité des répondants, des estimations découlant de la surveillance et du nombre d'interviews menées, ce qui pourrait avoir une incidence sur d'autres décisions opérationnelles. Actuellement, le groupe des répondants tardifs est petit et pourrait ne pas avoir d'effet sur les estimations découlant de la surveillance. Toutefois, la tendance actuelle des réponses aux enquêtes laisse présager une baisse continue du nombre de répondants à l'enquête, ce qui pourrait prolonger le temps nécessaires pour joindre le répondant et faire augmenter éventuellement le nombre proportionnel de répondants tardifs. Dans ces conditions, un examen minutieux de l'incidence des répondants tardifs sur les estimations découlant de la

population locale et l'empathie pour son mode de vie particulier avant de commencer l'enquête proprement dite a été un facteur clé du succès des travaux effectués dans cette région. Brièvement, nous pouvons conclure que, malgré un certain nombre de difficultés, la collecte de données quantitatives dans les régions rurales d'Afrique subsaharienne est une tâche qui peut être accomplie en obtenant des résultats satisfaisants. Un plan d'enquête et des méthodes d'interview appropriés, élaborés en collaboration avec le personnel et les experts locaux peuvent permettre d'obtenir des données de qualité suffisante pour une analyse économique de la pauvreté et de la vulnérabilité.

## Remerciements

Nous remercions le Ministère fédéral de la coopération et du développement de l'Allemagne pour son appui financier tout au long de ce projet visant à alléger la pauvreté et à soutenir la sécurité alimentaire en améliorant la mise en valeur et la gestion des pêcheries de rivières en Afrique. Ce projet a été coordonné par WorldFish Center. Les auteurs remercient aussi les deux examinateurs anonymes pour la révision complète de l'article et pour leurs précieux commentaires. Nos remerciements s'adressent aussi à l'éditeur pour nous avoir guidés au cours du processus de révision ainsi que pour d'autres conseils utiles apportés. Les idées exprimées dans l'article ne sont né- cessairement partagées par l'organisme subventionnaire ni par WorldFish Center.

## Bibliographie

- Béné, C., Neiland, A., Jolley, T., Ovie, S., Sule, O., Ladu, B., Babu, M., Mindjimba, K., Belal, E., Tiotsop, F., Dara, L., Zakara, A. et Quensiere, J. (2003a). Inland fisheries, poverty, and rural livelihoods in the Lake Chad Basin. *Journal of Asian and African Studies*, 38, 1, 17-51.
- Béné, C., Neiland, A., Jolley, T., Ovie, S., Sule, O., Ladu, B., Mindjimba, K., Belal, E., Tiotsop, F., Baba, M., Dara, L., Zakara, A. et Quensiere, J. (2003b). Natural-resource institutions and property rights in inland African fisheries - The case of the Lake Chad Basin region. *International Journal of Social Economics*, 30, 3, 275-301.
- Béné, C., Mindjimba, K., Belal, E., Jolley, T. et Neiland, A. (2003c). Inland fisheries, tenure systems and livelihood diversification in Africa: The case of the Yadeé floodplains in Lake Chad Basin. *African Studies*, 62, 2, 187-212.
- FAO (2005). Technical guidelines for responsible fisheries Nr 10: Increasing the contribution of small-scale fisheries to poverty alleviation and food security. FAO, Rome.
- FAO (2006). FAO's Activities on Small-scale Fisheries: An Overview. Advisory Committee on Fisheries Research (ACFR), Sixième Session, Rome, 17 au 30 octobre 2006.
- Jäckle, A., et Lynn, P. (2008). Offre de primes d'encouragement aux répondants dans une enquête par panel multimodales : effets cumulatifs sur la non-réponse et le biais. *Techniques d'enquête*, 34, 1, 115-130.
- Laaksonen, S. (2007). Pondération de données d'enquête recueillies en deux phases. *Techniques d'enquête*, 33, 2, 137-147.
- Neiland, A.E., Madaka, S.P. et Béné, C. (2005). Traditional Fisheries of Northern Nigeria. *Journal of Agrarian Change*, 5, 117-48.
- Neiland, A.E., Jaffry, S. et Kudasi, D.K. (2000). *Fishing Income, Poverty and Fisheries Management in North-East Nigeria*. Fisheries of North East Nigeria and the Lake Chad Basin, Volume I - A compilation of research project reports in two volumes 1993-2004, compilé par Dr. Arthur E. Neiland, 291-319.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- Rao, J.N.K. (2005). Evaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage. *Techniques d'enquête*, 31, 2, 127-151.
- Sarch, M.-T. (1997). Fishing and Farming in Lake Chad: Implications for Fisheries Development. *Development Policy Review*, 15, 141-57.
- UNEP (2004). Map of the Lake Chad Basin. Dans *Lake Chad Basin, GWA Regional assessment* 43, (Eds., M.P. Fortnam et J.A. Oguniola), University of Kalmar, Kalmar, Suède.



répondants puissent hésiter à fournir des renseignements à des représentants du gouvernement, un facteur plus important était que l'équipe de l'enquête représentait les deux groupes ethniques habitant la région de l'étude. En outre, les recenseurs parlaient les langues de la région, connaissaient bien les coutumes locales et étaient habitués aux conditions sur le terrain. De plus, la promesse d'un appui gouvernemental de suivi augmentait effectivement la disposition des répondants à fournir l'information.

Un autre atout des recenseurs choisis était leur sensibilité aux tensions ethniques. Ils ont fait très attention de ne prendre parti pour aucune des factions concernées et ont évité les phrases insultantes. Cet aspect était particulièrement important étant donné les multiples visites faites aux villages et aux répondants durant les enquêtes de suivi. Tout désaccord entre les répondants et les recenseurs aurait entraîné une érosion importante de l'échantillon et la nécessité de retirer des villages entiers de l'échantillon.

Certaines normes culturelles ou religieuses demandaient également du tact et du respect. Par exemple, dans plusieurs villages, seuls les hommes pouvaient être interviewés, parce que, dans cette culture africaino-musulmane, les femmes n'ont pas le droit d'être en contact ou de parler avec d'autres hommes que les membres de leur famille directe. Quand le chef de ménage n'était pas présent au moment de la visite, il n'était pas possible d'interviewer son épouse (ou toute autre femme faisant partie du ménage) à sa place. Un homme adulte faisant partie du ménage devait être choisi pour fournir l'information requise. Pour la même raison, les interviews ne pouvaient pas se dérouler dans la maison des répondants. Afin de se conformer à ces normes culturelles, il a fallu adapter la procédure d'interview. Au lieu de rendre visite aux ménages sélectionnés un par un, dans chaque village, les représentants de tous les ménages échantillonnés étaient convoqués par le chef du village dans un lieu de réunion central (habituellement devant la maison du chef). Si le chef du ménage n'était pas présent, un autre membre adulte du ménage (habituellement un homme) était interviewé. Les recenseurs s'asseyaient alors à trois à cinq mètres de distance l'un de l'autre et appelaient le répondant qui leur était assigné afin de l'interviewer en privé, tandis que les autres attendaient leur tour.

## Erosion de l'échantillon

Un défi particulier des enquêtes par panel consiste généralement à maintenir la taille de l'échantillon au cours du temps (Jackle et Lynn 2008, Laaksonen 2007). L'érosion peut être importante pour plusieurs raisons. Ainsi, dans certains cas, le chef du ménage était décédé, le ménage entier avait déménagé, ou les répondants n'étaient plus intéressés à participer, surtout si on ne leur offrait pas de prime d'incitation ou si celle-ci n'était pas suffisante. La

perte du désir de participer à une enquête de suivi a causé un problème durant la deuxième visite. À cause de contraintes budgétaires, l'équipe de l'enquête a décidé de ne pas dédommager les participants de leur temps à la seconde visite. Pour l'enquête de référence, chaque répondant avait reçu une boîte de sucre et un paquet de thé qui s'étaient avérés être un puissant incitatif extrinsèque. Quand les ménages ont appris qu'aucune rémunération n'était prévue à la deuxième visite, 69 d'entre eux (23 % du total de l'échantillon) ont annoncé qu'ils étaient « trop occupés » pour participer. Étant donné cette réaction, une rémunération à de nouveau été offerte à la troisième enquête, afin de pouvoir récupérer la plupart des ménages perdus au suivi. Ils étaient même disposés à répondre aux deux questionnaires (1<sup>er</sup> et 2<sup>e</sup> suivis). Par conséquent, les données manquantes ont pu être obtenues durant la dernière enquête, quoiqu'au prix d'une moins grande fiabilité à cause du biais de remémoration. Ce genre de comportement des répondants corrobore les résultats de Jackle et Lynn (2008), selon lesquels continuer d'offrir un paiement incitatif a des effets positifs significatifs sur l'érosion, le biais et la non-réponse partielle. À la fin de la période d'enquête, 14 ménages (4,7 %) avaient été perdus au suivi à cause d'une migration permanente ou d'autres raisons, et ont donc été supprimés de l'échantillon.

## 5. Résumé et conclusions

La collecte de données en vue d'analyser la pauvreté en Afrique subsaharienne est une tâche difficile. Souvent, des obstacles culturels, écologiques et économiques obligent les chercheurs à accepter un compromis entre la qualité des données et la faisabilité de l'étude. Par ailleurs, la collecte de ce genre de données est importante, parce que l'on en sait peu sur la pauvreté et la vulnérabilité des groupes marginalisés, tels que les communautés de pêcheurs dans les régions éloignées de l'Afrique subsaharienne. Dans le présent article, nous présentons l'approche qui a été adoptée au cours d'une étude sur la pauvreté et la vulnérabilité dans la plaine d'inondation du Logone, qui est une grande zone de pêche dans le Nord du Cameroun. Nous cernons les obstacles caractéristiques qui entravent souvent les travaux empiriques en Afrique subsaharienne et montrons comment ils peuvent être contournés grâce à un plan d'enquête et un échantillonnage adéquats et une application prudente de l'instrument d'enquête. Les principaux obstacles ont été la difficulté d'accès aux populations cibles, les difficultés à trouver des recenseurs qualifiés et la sensibilité culturelle importante requise de la part de l'équipe de chercheurs. Il est de la plus haute importance de collaborer étroitement avec les autorités locales et les experts des domaines de recherche, et de bien comprendre et respecter les normes et valeurs culturelles locales. L'apprentissage auprès de la



fin de février, l'accès aux villages échantillonnés était tout à fait impossible. L'équipe de recherche a opté pour un compromis, c'est-à-dire recueillir les données en décembre, même si cette période tombe au milieu de la saison des récoltes. Les données manquantes sur les rendements et le revenu ont ensuite été de nouveau recueillies durant la deuxième enquête de suivi. Des problèmes semblables se posent dans d'autres grandes régions de pêche continentale, telles que les zones humides de Hadejia-Nguru au Nigeria ou le bassin inférieur de la rivière Shire au Malawi.

#### *Définition des périodes d'enquête*

Dans le cas des enquêtes basées sur la remémoration, en particulier les enquêtes par panel (c'est-à-dire des enquêtes durant lesquelles l'équipe de recherche rend visite à plusieurs reprises aux mêmes ménages), il est important de s'assurer de la compréhension commune de la période qui est visée par le questionnaire. Diverses notions de durée peuvent donner lieu à des données entachées d'un biais sur les flux de revenus et de consommation et fausser les résultats et les conclusions tirés de l'étude. Afin d'être certains que la compréhension de la période pour laquelle les données sont demandées est uniforme, il faut tenir compte des différences culturelles de compréhension du temps. Nous avons constaté que, dans la plaine d'inondation du Logone, les habitants ne pensent pas en unités de temps comme les semaines ou les mois. Donc, des questions telles que « Combien d'argent avez-vous consacré à l'achat d'aliments au cours des six derniers mois? » ne convenaient pas. Dans ce cas particulier, il s'est avéré efficace de faire référence à certains événements sociaux ou festifs reconnus à l'échelle de la région. Par exemple, l'enquête de novembre coïncidait avec la fête du mouton (Tabaski), de sorte qu'il a été facile pour les répondants de délimiter la période prise en considération dans la deuxième enquête de suivi.

#### *Choix des recenseurs et de leurs connaissances culturelles*

L'élément qui est peut-être le plus important dans les travaux empiriques est le choix des recenseurs. Afin d'obtenir des données de bonne qualité, ces derniers doivent non seulement posséder les compétences et les connaissances nécessaires, mais également faire preuve de compétences générales supplémentaires, dont la maîtrise des langues, le savoir-faire dans les relations sociales et la volonté de travailler dans des conditions difficiles.

La pénurie d'intervieweurs suffisamment instruits dans la région de l'Extrême-Nord du Cameroun constituait une contrainte sérieuse. Pour les besoins de l'étude, cinq employés du MINEPPIA, qui travaillaient en tant que représentants gouvernementaux dans le domaine des enquêtes, ont été recrutés pour former l'équipe de recenseurs. Bien que les

consacrée aux groupes de discussion. Le chef convoquait alors les chefs des ménages sélectionnés dans un lieu de réunion central, habituellement sous un arbre en face de sa maison. Après l'interview, qui durait ordinairement environ une heure, un petit cadeau était remis au répondant pour le dédommager de son temps (un paquet de sucre et un sac de thé), et le chef du ménage suivant était invité à venir s'asseoir. Le travail en groupe permettait à l'équipe de terminer les interviews dans un village en un jour ou deux et de passer au village suivant. Cette façon de procéder a motivé et encouragé fortement les recenseurs, pour des raisons de sécurité et des raisons psychologiques. La durée de l'interview et, donc, le temps qu'il était prévu de passer dans un village ont été maintenus flexibles, afin de pouvoir procéder à une contre-vérification prudente de la cohérence et de la plausibilité des réponses. Par conséquent, durant les ateliers de formation des recenseurs et tout au long de la collecte des données, une importance particulière a été accordée à la prépondérance ultime de la qualité des données.

### **4. Défis de la collecte des données et leçons apprises**

La présente section décrit certains défis de la collecte des données et certains obstacles qu'il a fallu surmonter durant la présente étude, mais qui ne sont pas uniques à la région étudiée. Des conditions semblables s'observent dans de nombreuses zones humides et plaines d'inondation de l'Afrique subsaharienne, et les leçons tirées de cette étude pourraient s'avérer utiles pour des projets de collecte de données comparables.

#### *Saisonnalité*

Lorsque l'on recueille des données dans les communautés rurales dépendantes de la pêche de l'Afrique subsaharienne, la nature saisonnière des systèmes de subsistance et les contraintes écologiques doivent être prises en considération. Très souvent, les villages sont spatialement marginalisés et l'accès est extrêmement difficile durant certaines périodes de l'année. Ainsi, dans la plaine d'inondation du fleuve Logone dans le Nord du Cameroun, deux fois par an, l'accès aux villages est très restreint pendant plusieurs semaines à cause du cycle annuel d'inondations. Au début de la saison des inondations, et durant la période de désinondation, l'accès n'est possible ni par la route, ni par bateau. Donc, le choix des périodes d'enquête doit être adapté à ces conditions. Par exemple, alors qu'il aurait été plus raisonnable d'effectuer une enquête de suivi à la fin du cycle de production en janvier, afin de mieux refléter la production agricole et les prises de poissons, cette procédure s'est avérée irréalisable. De la mi-décembre à la

[Élevage, de la Pêche et des Industries Animales,

MINEPIA).

Tous les villages sélectionnés ont été visités avant le

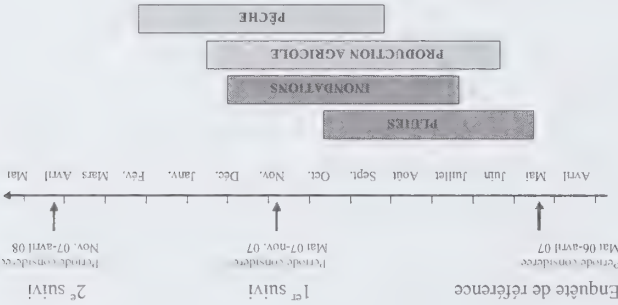
début de l'enquête auprès des ménages en vue d'établir un contact entre le chercheur et le chef du village, et de procéder à des groupes de discussion avec les dirigeants villageois. L'objectif des groupes de discussion était double. Premièrement, certains renseignements généraux ont été recueillis, tels que la taille et l'infrastructure du village, et l'accès aux ressources poissonnières et aux marchés du poisson. Deuxièmement, pour chaque village sélectionné, la liste complète des ménages a été dressée, puisqu'aucune

statistique officielle n'existait. Aux fins de l'étude, un ménage a été défini comme une unité économiquement indépendante constituée du chef du ménage, d'une ou de plusieurs épouses, d'enfants et d'autres membres directe- ment à charge, vivant dans le ménage ou ayant migré dans d'autres endroits. La taille du ménage varie de 2 personnes (c'est-à-dire normalement le mari et son épouse) à plus de 15. Les grands ménages sont fréquents dans le Nord du Cameroun, où la polygamie est très répandue, car les chefs de ménage vivent souvent avec jusqu'à quatre épouses. Le plus souvent, au lieu de vivre séparés, les ménages appa- rentés forment habituellement un clan et vivent ensemble dans une grande enceinte. Cependant, dans l'enceinte, les ménages sont indépendants les uns des autres. Durant les visites, une attention particulière a été accordée à la liste des noms des chefs de ménage individuels, plutôt qu'unique- ment à celle des dirigeants du clan. L'information supplé- mentaire recueillie durant les groupes de discussion a été nécessaire pour se faire une première idée des options et contraintes de subsistance dans la région étudiée, ce qui s'est avéré utile pour l'élaboration du questionnaire des ménages. À la dernière étape, les listes de ménages pro- duites ont été utilisées pour procéder à l'échantillonnage aléatoire pondéré des 300 ménages de l'échantillon.

### 3. Plan de sondage

La saisonnalité est une caractéristique importante des conditions de subsistance dans la plaine d'inondation du Logone. Par conséquent, afin de refléter les variations saisonnières, l'enquête a été conçue en vue de produire un ensemble de données de panel sur deux périodes (2006-2007), avec réalisation d'une troisième enquête six mois après l'enquête de référence (voir la figure 1). L'enquête de référence a été effectuée juste à la fin de la saison sèche, quand les activités génératrices de revenus étaient très limitées et que les ressources financières produites durant la saison des pluies de 2006 étaient en cours d'épuisement. L'enquête de référence, dont la période s'étend de mai 2006 à avril 2007, constitue une vérification des flux moyens de

revenus, de dépenses de consommation et du stock d'actifs. La première enquête de suivi a été menée durant la période occupée de l'année, où les dépenses augmentent à cause des investissements (par exemple achats de nouveaux filets de pêche et autres actifs productifs), et des coûts de production variables de l'agriculture et de la pêche. Enfin, la deuxième enquête de suivi couvrait la deuxième moitié de l'année, rendant compte des activités économiques des ménages durant cette période. Cette approche a été choisie en vue d'améliorer l'exactitude des données sur les activités de subsistance en raccourcissant la période de remémoration et de s'assurer de refléter la variation saisonnière du revenu et de la consommation.



Source : Illustration des auteurs

Figure 1 Options de subsistance dans la région étudiée et plan de l'enquête

Avant le début de chaque enquête, nous avons organisé des ateliers de formation des recenseurs d'une durée de trois à quatre jours durant lesquels a été effectué le prétest du questionnaire afin de déceler les faiblesses et la nécessité éventuelle d'éliminer, de reformuler ou d'ajouter des questions. Le prétest de l'enquête de référence a été exécuté dans deux villages des zones 1 et 2, afin de déterminer si le questionnaire convenait aux divers modes de subsistance. L'étude de référence a été achevée en trois semaines, en mai 2007, par quatre recenseurs travaillant en équipe, qui ont été supervisés directement par le premier auteur. Cette procédure a donné l'occasion d'effectuer une vérification croisée immédiate en vue de déceler l'information manquante, et a permis aux chercheurs d'observer et de renforcer les techniques d'interview et de discuter immédiatement des problèmes ou des questions.

Étant donné l'éloignement relatif des villages et les difficultés d'accès, la planification logistique a dû être minutieuse. Les déplacements sur les lieux duraient souvent plusieurs jours et il était inévitable de passer les nuits dans les villages. Donc, nous avons adopté la procédure d'enquête suivante. L'équipe complète arrivait dans un village, se présentait au chef du village, qui avait été informé au préalable de la date d'arrivée de l'équipe durant la visite



du sous-système plus important du Logone et du Chari dans le bassin du lac Tchad. Ce sous-système fournit 95 % du total des apports fluviaux au lac Tchad et la superficie du bassin est d'environ 650 000 km<sup>2</sup> (UNEP 2004). À l'intérieur de cette vaste surface, nous avons défini une région représentative en collaboration avec des experts nationaux et d'autres sources clés, tout en tenant compte de l'accessibilité et de la faisabilité logistique de l'étude. La région de l'étude couvre une superficie d'environ 2 400 km<sup>2</sup>, qui s'étend du lac Mga dans le sud au village Ivyé dans le Nord, où le Logomaya se joint au fleuve Logone. Cette région, dont la densité de population est relativement forte, est caractérisée par de riches stocks de poissons et des activités intensives de pêche, de transformation du poisson et de négoce du poisson.

Les moyens de subsistance de la population rurale de cette région sont particulièrement sensibles aux rudes conditions climatiques, dont une pluviosité limitée et irrégulière qui cause d'importantes variations de la production d'année en année (à cet égard, la région de l'étude est représentative de nombreux milieux ruraux semblables, particulièrement dans la zone sudano-sahélienne de l'Afrique subsaharienne) et, par conséquent, un risque considérable en matière de revenus. Cependant, l'effet diffère dans les diverses sous-régions de la région étudiée. Selon Neyman (1938), tel qu'il est cité dans Rao (2005), une méthode d'échantillonnage aléatoire stratifié a donc été considérée comme étant la plus efficace. Afin de tirer un échantillon représentatif de ménages dans la région étudiée tout en tenant compte des différentes conditions de production (telles que l'accès aux stocks de poissons), nous avons procédé à une stratification de la région étudiée en diverses zones agroécologiques. Nous avons supposé que, sous diverses conditions écologiques et de production, le rôle de la pêche en tant que générateur de revenus varierait. Cette procédure nous a permis de refléter le spectre complet de l'intensité de la pêche (allant des pêcheurs spécialisés/à temps plein aux ménages vivant purement de l'agriculture/l'élevage).

Ensuite, nous avons dressé la liste complète des villages compris dans la région étudiée ( $N=88$ ). Ces villages ont servi d'unités primaires d'échantillonnage. En suivant les recommandations des experts locaux de la pêche, nous avons sélectionné 14 villages proportionnellement au nombre total de villages par zone. La taille moyenne d'un village dans la plaine d'inondation (région étudiée) est d'environ 45 ménages, la fourchette allant de 15 à 100 ménages. Dans chaque village sélectionné, nous avons choisi au hasard un ménage sur deux sur la liste dressée par le chef du village. D'où, un échantillon de 300 ménages a été sélectionné proportionnellement à la taille des populations des villages, ce qui représente une fraction d'échantillonnage de 7 % de la population totale (estimée à 20 000 par le ministère de

d'enquêtes socioéconomiques auprès des ménages. D'autres contraintes de la collecte des données sont d'origine culturelles, comme les tensions entre divers groupes ethniques, l'existence d'une multitude de langues et de patois parlés dans la région où est effectuée l'étude, ou certaines particularités de la culture africaino-musulmane.

L'obtention des données nécessaires à l'évaluation de la pauvreté et de la vulnérabilité requiert une méthodologie d'enquête appropriée, afin que la qualité des données satisfasse aux exigences d'une analyse économétrique robuste. Les besoins de données en vue d'évaluer la pauvreté économique et de déterminer la contribution de la pêche artisanale au soulagement de la pauvreté et de la vulnérabilité sont grands. Des renseignements détaillés sur le revenu du ménage, y compris les diverses sources de revenus, telles que la production agricole, la pêche, l'élevage, le travail hors ferme, etc., sont nécessaires. Il faut aussi recueillir des données sur le stock et la valeur des actifs productifs et convertibles, ainsi que sur la répartition des dépenses de consommation. En outre, des renseignements sont nécessaires sur des variables de contrôle, telles que les chocs écologiques, économiques ou sociaux survenus dans le passé, les évaluations subjectives du risque, les dettes et autres obligations, la composition du ménage, et d'autres.

Le présent article décrit la procédure de collecte de données quantitatives auprès des ménages pauvres dans la plaine d'inondation du fleuve Logone, une importante région de pêche continentale du Nord du Cameroun. L'objectif de la collecte de données de panel au niveau du ménage effectuée en 2007-2008 était d'évaluer le rôle de la pêche artisanale dans l'atténuation des risques grâce à la diversification du portefeuille d'activités, ce qui contribue à réduire la vulnérabilité à la pauvreté. Dans le présent article, nous nous concentrons sur les exigences liées à l'approche méthodologique générale de l'échantillonnage et du plan de sondage. La nature complexe du secteur de la pêche artisanale mentionnée plus haut rend nécessaire l'élaboration d'une méthode d'échantillonnage et de collecte de données permettant d'évaluer la pauvreté et la vulnérabilité des ménages qui tirent leur subsistance de la pêche artisanale. En particulier, le plan de sondage doit tenir compte de la forte variation des activités génératrices de revenus au cours du temps en raison de la grande variabilité de l'accès aux ressources naturelles et des ajustements résultants de la situation de sécurité alimentaire, de la consommation, du revenu et des actifs des ménages.

## 2. Lieu de l'étude et méthode d'échantillonnage

Le lieu de l'étude est la plaine d'inondation du fleuve Logone dans la région de l'Extrême-Nord du Cameroun. La



# Collecte de données pour évaluer la pauvreté et la vulnérabilité dans les régions éloignées en Afrique subsaharienne

Rudolf Witt, Diemuth E. Pemsl et Hermann Waibel<sup>1</sup>

## Résumé

La collecte de données en vue d'évaluer la pauvreté en Afrique prend du temps, est coûteuse et peut présenter de nombreux obstacles. Dans le présent article, nous décrivons une procédure de collecte des données auprès de ménages vivant de la pêche continentale artisanale, ainsi que d'activités agricoles. Un plan d'échantillonnage a été établi afin de tenir compte de l'hétérogénéité des conditions écologiques et de la saisonnalité des moyens de subsistance possibles. Ce plan d'échantillonnage comprend une enquête par panel en trois points auprès de 300 ménages. Les répondants appartiennent à quatre groupes ethniques distincts sélectionnés aléatoirement parmi trois strates, chacune représentant une zone écologique différente. La première partie de l'article donne des renseignements contextuels sur les objectifs de la recherche, le lieu de l'étude et le plan de sondage, qui ont orienté le processus de collecte des données. La deuxième partie de l'article offre une discussion des obstacles qui entravent habituellement les travaux empiriques en Afrique subsaharienne et montre comment divers problèmes ont été résolus. Ces leçons pourraient aider les chercheurs à concevoir des enquêtes socioéconomiques appropriées dans des conditions comparables.

Mots clés : Enquêtes socioéconomiques auprès des ménages ; plan de sondage ; défis de la collecte des données ; Afrique subsaharienne.

## 1. Introduction

La collecte de données économiques sur la pêche artisanale en Afrique subsaharienne pose des défis, car les profils et les contraintes d'utilisation des ressources varient considérablement, dans l'espace, dans le temps et selon la saison. Par conséquent, la collecte des données nécessaires pour effectuer une évaluation significative de la pauvreté et de la vulnérabilité doit être planifiée minutieusement. Bien que la pêche artisanale puisse générer des bénéfices importants et contribuer considérablement au soulagement de la pauvreté et à la sécurité alimentaire, on possède peu de renseignements sur sa contribution réelle à la subsistance et à l'économie des ménages en Afrique subsaharienne (FAO 2005, 2006). Les principales entraves à la réalisation d'études empiriques dans ce domaine sont les difficultés associées à la collecte des données, comme l'éloignement et l'inaccessibilité des populations, surtout pendant la saison des pluies. La grande variabilité des conditions relatives aux ressources naturelles, et donc à la production, accroît les exigences ayant trait au plan de sondage. Pour préparer et mettre en œuvre une enquête en Afrique subsaharienne, les chercheurs peuvent s'inspirer de la méthodologie d'enquête, de la conception du questionnaire et des méthodes d'interview d'études comparables effectuées dans d'autres parties du monde, un exemple étant le questionnaire de la Living Standard Measurement Survey (LSMS) de la Banque mondiale. Cependant, les nombreuses particularités des communautés rurales de l'Afrique subsaharienne rendent nécessaire une approche adaptée et élaborée.

Certains de ces particularités sont de nature écologique, comme les variations saisonnières de l'accès aux ressources et aux marchés, lesquelles ont une incidence directe sur les profils et contraintes d'utilisation des ressources. D'autres ont trait à l'aspect économique du comportement des ménages, puisque les combinaisons d'activités génératrices de revenus des ménages ruraux de l'Afrique subsaharienne sont complexes. En particulier, les ménages appartenant aux communautés qui dépendent de la pêche ont adopté une matrice flexible et fortement saisonnière d'activités diversifiées (Béné, Neiland, Jolley, Ovie, Sule, Ladu, Mindjimba, Belal, Tiotso, Dara, Zakara et Quensiere 2003a ; Béné, Neiland, Jolley, Ladu, Ovie, Sule, Baba, Belal, Mindjimba, Tiotso, Dara, Zakara et Quensiere 2003b ; Béné, Mindjimba, Belal, Jolley et Neiland 2003c ; Neiland, Jaffry et Kudasi 2000, Neiland, Madaka et Béné 2005 ; Sarch 1997). Les populations locales sont alternativement ou simultanément pêcheurs, gardiens de troupeaux et agriculteurs, et chaque parcelles de terre peut être un lieu de pêche, un pâturage ou un champ cultivé, selon la phase du cycle des inondations (Béné et coll. 2003a, page 20). Étant donné la grande vulnérabilité du système écologique et économique aux chocs, tels que les inondations, la sécheresse et les infestations par des ravageurs, qui entraînent des variations annuelles des stocks de poissons et d'importantes pertes de récoltes, les ménages ont diversifié leur portefeuille d'activités, afin de répartir le risque de perte de revenus. Bien saisir l'interaction dynamique des divers modes de subsistance représente un défi particulier dans l'exécution

1. Rudolf Witt et Hermann Waibel, Institute of Development and Agricultural Economics, Faculty of Economics and Management, Leibniz University Hannover, Königsworther Platz 1, 30167 Hannover, Allemagne. Courriel : witt@ifgeb.uni-hannover.de ; Diemuth E. Pemsl, économiste, The Worldfish Centre, Penang, Malaisie.

## 7. Résumé

Dans le présent article, nous avons utilisé certains éléments théoriques de l'échantillonnage indirect et de l'échantillonnage en réseau pour démontrer un cadre statistique pour la conception et l'analyse d'enquêtes par téléphone mobile. Nous avons présenté un estimateur sans biais du total de population en ce qui concerne les unités d'estimation liées aux unités d'échantillonnage. Par implication, cette théorie offre un moyen de construire des estimateurs pour d'autres paramètres de population qui peuvent être exprimés sous forme de fonction de totaux. Nous avons illustré les problèmes en prenant comme exemple la NIS, qui est une enquête téléphonique portant sur les jeunes enfants et les adolescents.

Des renseignements recueillis durant l'interview de l'enquête sont nécessaires pour classer les unités d'estimation dans le domaine des utilisateurs de téléphones mobiles seulement, le domaine des utilisateurs de téléphones fixes seulement ou le domaine mixte. L'erreur de déclaration pourrait donner lieu à des erreurs de classification et nuire à l'absence de biais de l'estimateur, comme le pourrait la non-réponse à l'enquête dans les interviews par téléphone mobile et par téléphone fixe.

## Remerciements

Les auteurs remercient le rédacteur en chef adjoint pour ses précieux commentaires.

## Bibliographie

Arthur, A. (2007). The birth of a cellular nation. *The Source*. Mediamark Research Inc. Disponible au [http://www.mediamark.com/mri/TheSource/sorc2007\\_09.htm](http://www.mediamark.com/mri/TheSource/sorc2007_09.htm), 3.

Blumberg, S.J., et Luke, J.W. (2008). Wireless substitution: Early release of estimates from the National Health Interview Survey. National Center for Health Statistics. Disponible au <http://www.cdc.gov/nchs/nhis.htm>.

Brick, J.M., Dipko, S., Presser, S., Tucker, C. et Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.

Brick, J.M., Edwards, W.S. et Lee, S. (2007). Sampling telephone numbers and adults, interview length, and weighting in the California Health Interview Survey cell phone pilot study. *Public Opinion Quarterly*, 71, 793-813.

Cantor, J., Brownlee, S., Zukin, C. et Boyle, J. (2008). Do We Need to Worry About Wireless Substitution in Public Opinion Polls about Health Reform. Présentation au AcademyHealth 25<sup>th</sup> Annual Research Meeting, Washington, DC.

Carley-Baxter, L., Peytchev, A. et Lynberg, M. (2008). Comparison of cell phone and landline surveys: A design perspective. Document présenté au Annual meeting of the American Association for Public Opinion Research, New Orleans, LA.

Cassel, C.-M., Särndal, C.-E. et Wirtman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chowdhury, S., Montgomery, R. et Smith, P.J. (2008). Adjustment for noncoverage of nonlandline telephone households in RDD Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

CTIA (2008). Wireless Quick Facts. Disponible au <http://www.ctia.org/advocacy/research/index.cfm/AID/10323>.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, 32, 185-196.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Ehlen, J., et Ehlen, P. (2007). Cellular-only substitution in the United States as lifestyle adoption: Implications for telephone survey coverage. *Public Opinion Quarterly*, 71, 717-733.

Frankel, M., Battaglia, M., Link, M. et Mokdad, A. (2007). Integrating cell phone numbers into Random Digit-Dialed (RDD) landline surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, (Alexandria, VA), 3793-3800.

Harley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Keeter, S. (1995). Estimating non-coverage bias from a phone survey. *Public Opinion Quarterly*, 59, 196-217.

Khare, M., Singleton, J.A., Wouhib, A. et Jain, N. (2008). Assessment of Potential Bias in the National Immunization Survey (NIS) from the Increasing Prevalence of Households Without Landline Telephones. Présentation au National Immunization Conference, Centers for Disease Control and Prevention.

Lavallée, P. (2007). *Indirect Sampling*. New York : Springer Science+Business Media, LLC.

Lavarakas, P.J., Shuttles, C.D., Steeh, C. et Fienberg, H. (2007). The state of surveying cell phone numbers in the United States: 2007 and Beyond. *Public Opinion Quarterly*, 71, 840-854.

Smith, P.J., Hoaglin, D.C., Battaglia, M.P., Khare, M. et Barker, L.E. (2005). Statistical methodology of the National Immunization Survey, 1994-2002. National Center for Health Statistics. Hyattsville, MD. *Vital and Health Statistics*, Séries 2, 138.

Wolter, K.M. (2007). *Introduction to Variance Estimation*, Second Edition. New York : Springer-Verlag.

Wolter, K.M., Chowdhury, S. et Kelly, J. (2008). Design, conduct, and analysis of random digit dialing surveys. Dans *Handbook of Statistics: Sample Surveys, Theory, Methods and Inference*, (Eds., D. Pfeffermann et C.R. Rao), Elsevier, Oxford, UK.



qu'environ 18 % des enfants admissibles et 10 % des adolescents admissibles pourraient ne pas figurer dans la base de sondage de la NIS. Pour faire face à l'accroissement du nombre de ménages dotés d'un téléphone mobile seulement dans la population cible de la NIS, des interviews par téléphone mobile pourraient être ajoutées à l'enquête.

Dans le cas de la NIS, le numéro de téléphone est l'UE, la mère ou le père bien informé est l'UR et l'enfant admissible est l'UE. Pour l'échantillon par CA ou échantillon A, le parent est un membre du ménage auquel le numéro de ligne fixe dans l'échantillon est attribué, tandis que pour l'échantillon de lignes de téléphone mobiles ou échantillon B, le parent possède l'accès régulier au téléphone mobile auquel le numéro de téléphone de l'échantillon est attribué. Dans la NIS, au lieu de sous-échantillonner les enfants, le parent le mieux informé fournit les renseignements pour tous les enfants d'âge admissible qui vivent dans son foyer (mais pas pour les enfants qui pourraient vivre ailleurs). Ces éléments du protocole d'enquête établissent les liens entre les UR et les UE et entre les UE et les UR.

Un plan de sondage exhaustif pour la NIS consiste à effectuer l'estimation en se basant sur les domaines non chevauchants et la décomposition (3). Autrement dit, l'échantillon A est utilisé pour représenter tous les enfants liés à un ménage possédant un téléphone fixe et l'échantillon B est utilisé pour représenter tous les enfants liés à un parent possédant un téléphone mobile seulement. Nous avons considéré et rejeté les décompositions (4) et (5) pour des raisons de coût et de risque de biais de non-réponse différentiel dans l'estimation pour la population mixte. Pour appliquer l'estimateur donné par (10), nous déterminons si l'enfant compris dans l'échantillon A vit dans un ménage doté d'un téléphone fixe seulement en utilisant les trois questions suivantes :

- A1. Maintenant, je vais vous poser quelques questions au sujet des téléphones mobiles dans votre ménage. En tout, combien de téléphones mobiles en état de fonctionnement avez-vous, ainsi que les membres de votre ménage, à votre disposition pour usage personnel ? Veuillez ne pas compter les téléphones mobiles qui sont utilisés exclusivement pour affaires.
- A2. Combien de [ces] téléphones mobiles les parents et gardiens de [ÉNUMÉRER TOUS LES ENFANTS ADMISSIBLES] utilisent-ils habituellement ?
- A3. Les appels téléphoniques que vous et les membres de votre famille recevez, sont-ils presque tous reçus sur des téléphones mobiles, ou certains sont-ils reçus sur des téléphones mobiles et certains sur des téléphones ordinaires ? (SI LE RÉPONDANT DEMANDE S'IL FAUT INCLURE LES APPELS

D'AFFAIRES : Veuillez ne pas inclure dans votre réponse les appels liés aux affaires).

Pour l'échantillon de lignes de téléphone mobiles ou échantillon B, nous déterminons si l'enfant fait partie d'un ménage à téléphone mobile seulement au moyen des deux questions suivantes.

- B1. Avez-vous une ligne de téléphone fixe dans votre ménage ? (INTERVIEWEUR : APPROFONDIS-SEZ SI OUI : Veuillez ne pas inclure les lignes réservées uniquement aux modems, les lignes réservées uniquement aux télécopieurs, les lignes utilisées uniquement pour un système de sécurité à domicile, les téléavertisseurs ou les téléphones mobiles).
- B2. En pensant uniquement au téléphone raccordé à la ligne fixe, et non à votre téléphone mobile, si ce téléphone sonnait et que quelqu'un se trouvait à la maison, dans les circonstances normales, quelle est la probabilité que l'on réponde à l'appel ? Diriez-vous très probable, assez probable, assez improbable ou tout à fait improbable ?

Nous utilisons la question B2, due à Cantor, Brownlee, Zukin et Boyle (2008), pour déterminer si la ligne de téléphone fixe est effectivement utilisée pour des communications vocales et, donc, si le répondant appartient au domaine *ab* ou *b*.

Également pour l'échantillon B, afin de déterminer le nombre de téléphones mobiles dans la population qui sont liés à un enfant admissible donné, nous utilisons les deux questions suivantes :

- B3. Maintenant, je vais vous poser quelques questions au sujet des téléphones mobiles dans votre ménage. En tout, combien de téléphones mobiles en état de fonctionnement avez-vous, ainsi que les membres de votre ménage, à votre disposition pour usage personnel ? Veuillez ne pas compter les téléphones mobiles qui sont utilisés exclusivement pour affaires, et veuillez inclure le numéro que nous avons appelé.
- B4. Combien de [ces] téléphones mobiles les parents et gardiens de [ÉNUMÉRER ENFANTS] utilisent-ils habituellement ? Veuillez inclure le numéro que nous avons appelé.

Les réponses aux questions A1 à A3 et B1 à B4 permettent de calculer les poids de sondage et d'appliquer l'estimateur sans biais du total de population donné par (10).



## 6. Exemple : la National Immunization Survey (NIS)

Afin d'illustrer l'information qui doit être recueillie durant l'interview de l'enquête, nous nous servons de la NIS, qui est une enquête réalisée auprès des parents d'enfants de 19 à 35 mois et d'adolescents de 13 à 17 ans parrainée par les Centers for Disease Control and Prevention (CDC) afin de surveiller les taux de couverture de la vaccination (c'est-à-dire la proportion d'enfants dont la vaccination est à jour en regard du calendrier recommandé de vaccination) aux États-Unis. Les données recueillies dans le cadre de la NIS le sont en deux phases, à savoir une enquête téléphonique par CA auprès des ménages munis d'un téléphone fixe qui ont de jeunes enfants ou des adolescents dans la tranche d'âge admissible, suivie d'une enquête par la poste auprès des prestataires de la vaccination des enfants d'âge admissible. Pour la phase de l'enquête téléphonique de la NIS, la base de sondage comprend l'ensemble des numéros de téléphone au États-Unis dans les banques non vides de 100 numéros de téléphone consécutifs (banques de type 1+). À l'heure actuelle, les numéros de téléphone mobile regroupés sous les indicatifs de central réservés à la téléphonie mobile ne sont pas inclus dans la base d'échantillonnage de la NIS. Quand un ménage comptant un enfant d'âge admissible est repéré durant l'enquête téléphonique, l'interview est effectuée auprès de l'adulte dans le ménage qui est reconnu comme étant la personne la mieux informée au sujet de la situation de vaccination de l'enfant (presque toujours la mère ou le père). Durant l'interview téléphonique, des données sont recueillies pour chaque enfant d'âge admissible présent dans le ménage, y compris les caractéristiques démographiques de l'enfant, les caractéristiques démographiques de la mère et les caractéristiques socioéconomiques du ménage où vit l'enfant. À la fin de l'interview téléphonique, l'intervieweur demande l'autorisation de communiquer avec le prestataire de la vaccination de l'enfant. Si l'autorisation est accordée, tous les prestataires de vaccination nommés par le répondant durant l'interview téléphonique sont contactés par la poste afin d'obtenir les antécédents de vaccination de l'enfant déclarés par le prestataire, qui sont utilisés dans l'analyse statistique pour évaluer la situation de vaccination. Smith, Hoaglin, Battaglia, Khare et Barker (2005) décrivent en détail les méthodes statistiques utilisées par l'équipe de la NIS.

Etant donné la croissance de la population ayant un téléphone mobile seulement, la proportion de la population cible de la NIS qui est couverte par la base de sondage des lignes de téléphone fixes a diminué ces dernières années. En utilisant les données de la *National Health Interview Survey*, Khare, Singleton, Wouhib et Jain (2008) estiment

effectués comme auparavant.

Alternativement, le méthodologiste de l'enquête pourrait demander une étape de randomisation réelle, qui nécessiterait que l'intervieweur dresse la liste des UR liées à l'UE et qu'il sélectionne une de ces UR au hasard, ou une étape de pseudorandomisation en utilisant la méthode du dernier anniversaire. De telles méthodes ne sont probablement pas applicables en ce moment, à cause de la difficulté à obtenir la collaboration des personnes sélectionnées pour les interviews par téléphone mobile.

### 5.6 Incidences sur la collecte des données

Certaines données doivent être recueillies durant l'interview afin de permettre le calcul des estimateurs dont nous discutons ici.

Pour appuyer l'utilisation de  $\delta_j$ , l'enquête par téléphone mobile doit fournir des renseignements permettant d'établir si une des UR liées à l'UE a accès à un téléphone fixe. L'UR répondante doit fournir cette information en ce qui la concerne ainsi que pour d'autres UR qui pourraient être liées à l'UE.

Afin d'appuyer l'utilisation de  $\phi_j$ , l'enquête par ligne téléphonique fixe doit fournir des renseignements permettant d'établir si l'une des UR liées à l'UE possède un accès régulier à un téléphone mobile. L'UR répondante doit fournir cette information en ce qui la concerne ainsi que pour d'autres UR qui pourraient être liées à l'UE. Cette déclaration peut être assez simple si le protocole de réponse relie uniquement les UE aux UR présentes dans le même ménage. Pour des protocoles de réponse plus compliqués, la déclaration pourrait être difficile à obtenir.

Afin d'appuyer l'utilisation de  $\sum_{i \in U_{EB}} \ell_{ij}$  dans le calcul des poids de sondage, des données doivent être recueillies durant l'enquête afin d'établir combien d'UE dans la population sont liées à l'UE  $j$  pour laquelle est faite la déclaration. L'UR répondante doit être capable de donner le nombre de téléphones mobiles, y compris le sien, dont les appels parviennent à une UR qui est liée à une UE donnée.

Si l'estimateur donné par (21) et (22) est utilisé afin d'identifier toutes les UE, des renseignements supplémentaires doivent être recueillis durant l'interview. L'UR répondante doit connaître et indiquer le nombre d'UR, y compris elle-même, qui sont liées à la fois à l'UE sélectionnée et à l'UE pour laquelle est faite la déclaration. L'UR répondante doit également connaître et indiquer sa part de l'utilisation du téléphone mobile sur lequel l'interview est effectuée ou être capable de préciser que l'utilisation est approximativement égale.

la variance de l'estimateur GREC dans un contexte d'échantillonnage indirect. Voir aussi Wolter (2007, chapitre 6) pour l'estimation de la variance de l'estimateur GREC.

Avant d'abandonner le sujet du calage, notons que nous

avons en grande partie laissé de côté la petite population n'ayant pas le téléphone, qu'il est fondamentalement impossible d'échantillonner dans une enquête téléphonique. Pourtant, selon toute vraisemblance, le total global de population  $Y^E = Y^{ET} + Y^{EC}$  sera le paramètre d'intérêt plutôt que le total de la population ayant le téléphone  $Y^{ET}$ , et les totaux de contrôle connus utilisés dans le calage pourraient être les totaux pour la population globale  $Z^E = Z^{ET} + Z^{EC}$ , et non les totaux pour la population ayant le téléphone  $Z^{ET}$ . Afin d'inclure la population sans téléphone, nous pourrions envisager d'utiliser un estimateur GREC révisé avec  $g_j = 1 + (Z^E - Z^{ET})/Z_j^E$ . Cette révision s'appuie sur le même modèle pour la population sans téléphone que pour la population ayant le téléphone. Voir Keeter (1995) et Chowdhury, Montgomery et Smith (2008) pour d'autres considérations concernant le calage des poids pour la population sans téléphone.

## 5.5 Hypothèses d'identifiabilité

La théorie qui précède repose sur l'hypothèse fondamentale que, si l'UE  $i$  est sélectionnée dans l'échantillon de lignes de téléphone mobiles, alors la variable  $X_i$  définie en (2) est observable. Néanmoins, le 9<sup>e</sup> réseau (comme le 8<sup>e</sup>) de la figure 1 illustrent un problème éventuel de cette théorie. Dans le cas de ce réseau, deux UR sont liées à une UE et, à leur tour, les UR sont liées chacune à une seule UE. Pour poursuivre cet exemple, nous supposons que ces deux UE ne sont liées à aucune autre UR dans la population. Au moment de l'interview de l'enquête, une seule des UR sera habituellement rejointe et interviewée (à moins que le protocole d'enquête exige spécifiquement qu'une tentative d'interview ait lieu auprès de chaque UR liée à l'UE sélectionnée). L'UR répondante fournira les renseignements pour l'UE à laquelle elle est liée, mais étant donné la nature même de ce réseau, le répondant ne peut pas faire de déclaration pour l'UE qui est liée à l'UR compagnon qui partage la ligne de téléphone mobile figurant dans l'échantillon. Donc, il existe au moins une UE liée à l'UE qui ne peut pas être observée, c'est-à-dire pour laquelle les données ne peuvent pas être recueillies durant l'interview par téléphone mobile. Donc, nous disons que  $X_i$  est non identifiable. La situation concernant la possibilité de déclaration pour les deux UE serait inversée si, au moment de la tentative d'interview par téléphone mobile, l'appel était reçu par l'UR compagnon.

Afin de préserver l'absence de biais de l'estimateur du total de population, la variable  $X_i$  doit être identifiable pour chaque UE répondante sélectionnée dans l'échantillon de

est donné par

$$X_i = \sum_{j \in U_{EB}} \frac{\delta_j Y_j \ell_{ij}}{\delta_j Y_j \ell_{ij}} = \sum_{j \in U_{EB}} \frac{\tau_{ik}}{\delta_j Y_j \ell_{ij}} \alpha_{ik} \quad (20)$$

Deuxièmement, une approche plus réaliste consisterait à émettre l'hypothèse d'une étape de randomisation supplémentaire, à savoir que la tentative d'appel pour l'interview devant être administrée à l'UE a atteint une UR liée à l'UE sélectionnée aléatoirement. Cette randomisation pourrait être considérée comme étant conceptuelle (autrement dit, survenant naturellement et non dirigée par le méthodologiste de l'enquête). De manière formelle et rigoureuse, nous devrions recueillir l'information sur le nombre d'UR liées à l'UE et la probabilité que la tentative d'appel sur le téléphone mobile soit reçue par l'UR répondante. La probabilité serait approximée par la part d'utilisation du téléphone mobile auto-déclarée par le répondant. Si une UR seulement est liée à l'UE, cette probabilité est de 1.0 et, manifestement, cette valeur simple ne devra pas être recueillie durant l'interview une fois qu'il a été signalé qu'il n'existe qu'une UR. Si deux UR ou plus sont liées à l'UE, la probabilité ou la part dont la valeur doit être recueillie est désignée par  $\tau_{ik}$  pour les UR portant l'indice inférieur  $k$ , où  $\sum_{k \in U_{RB}} \tau_{ik} = 1$  et  $U_{RB}$  est l'ensemble d'UR qui sont liées à la  $i^e$  UE. En disposant de cette information supplémentaire, un estimateur sans biais de

$$X_i = \sum_{j \in U_{EB}} \frac{\delta_j Y_j \ell_{ij}}{\delta_j Y_j \ell_{ij}} = \sum_{j \in U_{EB}} \frac{\tau_{ik}}{\delta_j Y_j \ell_{ij}} \alpha_{ik} \quad (20)$$

où  $\alpha_{ik}$  est une variable indicatrice signalant si la  $k^e$  UR est le répondant réalisé ou non pour la  $i^e$  UE dans  $s_{EB}$  et  $\ell_{ikj} = 1$ , si l'UE  $i$  est liée à l'UR  $k$  qui, à son tour, est liée à l'UE  $j$ ; autrement,

$$Y^{EB} = \sum_{j \in U_{EB}} \delta_j Y_j W_{0j}^{EB} \quad (21)$$

Les données sont maintenant identifiées et nous pouvons insérer (20) dans (7), ce qui donne l'estimateur révisé

$$W_{0j}^{EB} = \sum_{i \in s_{EB}} W_{EB}^i \frac{\sum_{k \in U_{RB}} \ell_{ikj}}{1} \alpha_{ik} \frac{\tau_{ik}}{\delta_j Y_j \ell_{ikj}} \quad (22)$$

avec les poids révisés

En guise d'approximation, nous pourrions supposer que les UR sont égales aux utilisateurs de téléphone mobile,



dans une enquête par téléphone mobile ; elles peuvent être spécifiées en fonction de l'information disponible à l'étape de pondération précédente ou de toute information recueillie durant l'interview de présélection. Le taux pondéré d'achèvement de l'interview est

$$R_{3j} = \frac{\sum_{i \in s_{jB}^*} r_{3i} e_{2i} W_{3i}^{EB}}{\sum_{i \in s_{jB}^*} e_{2i} W_{3i}^{EB}}.$$

Le total estimé pour le domaine des unités ayant un téléphone mobile seulement peut maintenant s'exprimer par

$$Y^{Eh} = \sum_{j \in s_{EB}} \delta_j X_j W_{4j}^{EB}, \quad (17)$$

où

$$W_{4j}^{EB} = \sum_{i \in s_{jB}^*} W_{4i}^{EB} \ell_{ij} / \sum_{i \in s_{jB}^*} \ell_{ij}$$

et  $s_{EB}^*$  est l'ensemble des UE admissibles déclarées durant les interviews de présélection. Le poids est nul pour toute UE admissible comprise dans  $s_{EB}^*$  pour laquelle l'UR n'a pas achevé de répondre à l'interview principale. Le total estimé pour le domaine des unités mixtes, s'il est requis en fonction du protocole d'enquête, est défini de la même façon par

$$Y^{Eab} = \lambda \sum_{j \in s_{EA}^*} W_{4j}^{EA} (1 - \phi_j) X_j + (1 - \lambda) \sum_{j \in s_{EB}^*} W_{4j}^{EB} (1 - \delta_j) X_j.$$

Pour le cas ii, la correction de la non-réponse à l'interview doit être faite au niveau de l'UE. Les UE sont traitées comme des cas engendrés et il faut décider pour chacune s'il s'agit d'un cas d'interview achevée. Pour le domaine des unités à téléphone mobile seulement, le total estimé est (17), où le poids est maintenant défini par

$$W_{4j}^{EB} = r_{3j} e_{2j} W_{3j}^{EB} / R_{3j} \quad \text{pour } j \in s_{EB},$$

$$W_{3j}^{EB} = \sum_{i \in s_{jB}^*} W_{3i}^{EB} \ell_{ij} / \sum_{i \in s_{jB}^*} \ell_{ij}$$

$$R_{3j} = \frac{\sum_{i \in s_{jB}^*} r_{3i} W_{3i}^{EB}}{\sum_{i \in s_{jB}^*} W_{3i}^{EB}}.$$

Ici, les cellules de pondération,  $r_j$ , sont définies en fonction des caractéristiques des UE, déterminées d'après l'interview de présélection ou d'autres sources.

Aussi bien pour le cas i que ii, afin de faciliter les calculs, posons que  $W_{4j}^{EA}$  est défini et égal à 0 pour les UE comprises dans l'échantillon de lignes de téléphone mobiles et

que les poids de sondage pour l'estimation du total de population d'enquête est celui de la section 5.1, nous concluons que l'échantillon de lignes de téléphone fixes. Si le protocole d'enquête est égal à 0 pour les UE comprises dans l'échantillon de lignes de téléphone fixes, nous concluons que les poids de sondage pour l'estimation du total de population d'intérêt sont définis par

$$W_j = W_{4j}^{EA} + W_{4j}^{EB} \delta_j \quad (18)$$

pour  $j \in s_{EB}^*$ , où  $s_{EB}^* \in s_{EA}^* \cup s_{EB}^*$ . Autrement, si le protocole d'enquête est celui de la section 5.2, nous concluons que les poids de sondage sont définis par

$$W_j = W_{4j}^{EA} \{\phi_j + \lambda(1 - \phi_j)\} + W_{4j}^{EB} \{\delta_j + (1 - \lambda)(1 - \delta_j)\} \quad (19)$$

pour  $j \in s_{EB}^*$ .

Les poids corrigés de la non-réponse donnés par (18) ou (19) peuvent être calés (Deville et Särndal 1992) sur des totaux de contrôle externes dans les cellules socioéconomiques ou géographiques pour la population d'UE, en appliquant des méthodes de poststratification, de calage (ratissage) ou d'estimation par la régression généralisée (GREG). Si des sources précises sont disponibles, les totaux de contrôle peuvent être établis et le calage peut être exécuté séparément pour les domaines  $A$  et  $B$  ou pour les domaines  $a, ab$  et  $b$ . Si l'on ne dispose pas de totaux de contrôle selon la situation concernant le téléphone, le calage doit se faire en utilisant des totaux de contrôle pour la population complète, indépendamment de la situation concernant le téléphone.

Afin d'illustrer ces idées, nous examinons brièvement l'estimateur GREG. Supposons que nous disposons d'une variable auxiliaire  $1 \times p, Z_j$ , pour les UE admissibles observées pour lesquelles les totaux de contrôle  $Z_{j \in s_{EB}^*}$  sont connus. Par exemple, la variable  $z$  peut être issue d'un modèle entièrement saturé en ce qui concerne les variables explicatives d'âge, de race et de sexe. Soit  $s_{EB}^*$  l'ensemble d'UE pour lesquelles l'interview principale est achevée et soit  $n_{EB}^* = \#(s_{EB}^*)$  le nombre d'UE admissibles déclarées durant les interviews achevées obtenues dans l'échantillon consolidé de lignes téléphoniques. Empilons les valeurs de  $y_j$ , les valeurs de  $z$  et les poids dans les matrices  $Y = (Y_1, \dots, Y_{n_{EB}^*})'$ ,  $Z = (Z_1, \dots, Z_{n_{EB}^*})'$  et  $W = \text{diag}(W_1, \dots, W_{n_{EB}^*})'$ . Alors, l'estimateur GREG (Cassel, Särndal et Wretman 1976) du total d'intérêt pour la population ayant le téléphone prend la forme familière

$$Y_{EB}^{ET} = Y_{EB}^{ET} + (Z_{EB}^{ET} - Z_{EB}^{ET}) \hat{\beta} = \sum_{j \in s_{EB}^*} W_j g_j Y_j,$$

où les coefficients estimés sont donnés par  $\hat{\beta} = (Z'WZ)^{-1}Z'WY$ ,  $Y_{EB}^{ET} = \sum_{j \in s_{EB}^*} W_j Y_j$ , et  $g_j = 1 + (Z_{EB}^{ET} - Z_{EB}^{ET})Z_j'$ ,  $Z_j' = (Z_j, \dots, Z_j)$  et  $g_j$  est défini et égal à 0 pour les UE comprises dans l'échantillon de lignes de téléphone mobiles et



données manquantes qui surviennent durant les interviews par téléphone mobile, en supposant que les corrections standard pour la présence de données manquantes dans l'échantillon de lignes de téléphone fixes ont déjà été intégrées dans les poids  $\{W_{EA}^f\}$ .

Les données manquantes peuvent être dues à trois facteurs, à savoir i) la non-résolution de l'UE, ii) une interview de présélection incomplète de l'UR et iii) une interview adoptions les conventions suivantes. L'étape de résolution fait référence à la classification de l'UE comme un NMPA ou autre chose, par exemple une ligne déconnectée ou une ligne réservée aux affaires. Les UE dont la situation n'est pas résolue et les UE classées comme n'étant pas un NMPA ne poursuivent pas l'interview. L'étape de la présélection fait référence à une brève interview préliminaire destinée à confirmer la situation concernant le téléphone et à déterminer toute caractéristique démographique ou autre caractéristique d'admissibilité de toute UE liée à l'UR; les UR pour lesquelles l'interview de présélection est incomplète ou pour lesquelles l'interview de présélection est incomplète ou

admissible n'est liée à l'UR ne passent pas à l'étape de l'interview. Si le protocole d'enquête requiert de n'inclure que les UE possédant un téléphone mobile seulement, comme à la section 5.1, l'interview se terminera à ce stade-la pour toute UE mixte. Par ailleurs, si le protocole d'enquête requiert d'inclure à la fois les UE à téléphone mobile seulement et les UE mixtes, comme à la section 5.2, alors l'interview se poursuivra pour toutes les UE de ce type. L'étape de l'interview fait référence à la collecte des réponses aux principaux items de l'enquête pour chaque UE admissible liée à l'UR. Le méthodologiste de l'enquête doit définir ce qui constitue une interview complète. En particulier, il doit décider si une *rupture* (une tentative d'interview qui est achevée pour certaines UE admissibles liées à l'UR, mais non toutes) doit être traitée comme une interview complète ou non. D'autres auteurs pourraient choisir d'organiser les étapes du processus de réponse à l'enquête d'une manière un peu différente de celle adoptée ici.

Les corrections des poids d'échantillonnage peuvent être faites pour la non-résolution et pour la non-réponse au questionnaire de présélection en supposant comme mécanisme de réponse que les données manquent au hasard. Ces deux corrections doivent être faites au niveau de l'UE. Soit  $\{s_{EB}^f\}$  une partition de l'échantillon de lignes de téléphone mobiles en cellules de pondération  $\alpha$  spécifiées par l'utilisateur et désignant maintenant les poids d'échantillonnage tirés de (6) par  $W_{EB}^{f_i}$ , où l'indice inférieur 1 a été ajouté simplement pour indiquer la première étape d'un processus de correction en plusieurs étapes. Les indicateurs téléphoniques régionaux et les variables de recensement géographiques au niveau de l'indicateur régional peuvent être

utilisées pour former les cellules de pondération; sinon, peuvent de la résolution particuliers aux téléphones mobiles sont définis par

$$R_{1\alpha} = \frac{\sum_{i \in s_{EB}^{f_i}} W_{EB}^{f_i}}{\sum_{i \in s_{EB}^{f_i}} W_{EB}^{f_i}},$$

où  $r_{1i}$  est une variable indicatrice de résolution (= 1, si le cas est résolu, = 0, s'il n'est pas résolu), et les poids corrigés de la non-résolution sont  $W_{EB}^{f_i} = r_{1i} W_{EB}^{f_i} / R_{1\alpha}$  pour  $i \in s_{EB}^{f_i}$ .

Soit  $e_{1i}$  une variable indiquant si  $i$  est un NMPA résolu (= 1, si NMPA résolu, = 0, autrement), et soit  $\{s_{EB}^{f_i}\}_{f_i=1}$  une partition de l'échantillon de lignes de téléphone mobiles en cellules de pondération spécifiées par l'utilisateur, qui pourraient être les mêmes que celles de la partition qui précède, ou être différentes. Alors, les taux d'achèvement de l'interview de présélection particuliers aux téléphones mobiles sont

$$R_{2\beta} = \frac{\sum_{i \in s_{EB}^{f_i}} r_{2i} e_{1i} W_{EB}^{f_i}}{\sum_{i \in s_{EB}^{f_i}} e_{1i} W_{EB}^{f_i}},$$

où  $r_{2i}$  est une variable indicatrice de présélection (= 1, si la présélection est achevée, = 0, si la présélection n'est pas achevée), et les poids corrigés de la non-réponse à la présélection sont  $W_{EB}^{f_i} = r_{2i} e_{1i} W_{EB}^{f_i} / R_{2\beta}$  pour  $i \in s_{EB}^{f_i}$ . Notons que la somme appropriée des poids est préservée à chaque étape du processus de correction.

Ensuite, les poids d'échantillonnage doivent être corrigés de la non-réponse à l'interview. Selon la classification adoptée pour les ruptures d'interview par le méthodologiste de l'enquête, deux cas doivent être pris en considération: i) l'UR achève ou n'achève pas l'interview pour l'ensemble des UE qui sont admissibles et qui lui sont liées, ou ii) l'UR achève ou n'achève pas l'interview de manière sélective sur la base des UE individuelles. Si les ruptures sont classées comme des interviews incomplètes, seul le cas i s'applique. Soit  $e_{2i}$  une variable indiquant si l'UR est présélectionnée et si elle est liée à au moins une UE qui est admissible à l'interview (= 1, si présélectionnée et admissible, = 0, autrement), et soit  $r_{3i}$  la variable indicatrice d'interview (= 1, si l'interview est complète, = 0, autrement).

Pour le cas i, la correction des poids peut être faite au niveau de l'UE et est donnée par  $W_{EB}^{f_i} = r_{3i} e_{2i} W_{EB}^{f_i} / R_{3\gamma}$  pour  $i \in s_{EB}^{f_i}$ , où  $R_{3\gamma}$  est le taux pondéré d'achèvement de l'interview calculé à l'intérieur des cellules de pondération  $\gamma$  spécifiées par l'utilisateur. De nouveau, les options pour la construction des cellules de pondération sont limitées

(2008). Comme l'échantillonnage est effectué indépendamment dans les bases de sondage de lignes fixes et de lignes mobiles, l'estimateur sans biais de la variance du total estimé pour la population complète ayant le téléphone devient

$$v(Y_{\text{ET}}^{\text{EB}}) = v(Y_{\text{EA}}^{\text{EB}}) + v(Y_{\text{EB}}^{\text{EB}}). \quad (15)$$

Afin de faciliter les développements qui suivent, soit  $V_{\text{EB}}[\delta Y]$  un autre symbole pour représenter l'estimateur de variance donné par (14). Cette notation mettra l'accent sur le fait que l'estimateur de variance est basé sur la variable  $X_i$  donnée par (13) définie en fonction de la caractéristique  $\delta_j Y_j$ , qui est la caractéristique d'intérêt pour les UE ayant un téléphone mobile seulement. En outre, posons que le symbole  $V_{\text{EA}}[Y]$  représente l'estimateur  $v(Y_{\text{EA}}^{\text{EB}})$  défini en fonction de la caractéristique  $Y_j$ . En nous servant de cette notation, (15) devient  $v(Y_{\text{ET}}^{\text{EB}}) = V_{\text{EA}}[Y] + V_{\text{EB}}[\delta Y]$ .

Deuxièmement, considérons l'estimation de la variance dans le cas de domaines chevauchants. L'estimateur du total pour la population ayant le téléphone est maintenant  $Y_{\text{ET}}^{\text{EB}}$  donné par (11). Pour une valeur fixe de  $\lambda$ , l'estimateur sans biais de la variance se dégage clairement du travail effectué dans (14) et (15). Il est donné par

$$v(Y_{\text{ET}}^{\text{EB}}) = V_{\text{EA}}[\phi Y + \lambda(1 - \phi)Y]$$

$$+ V_{\text{EB}}[\delta Y + (1 - \lambda)(1 - \delta)Y]. \quad (16)$$

Le premier terme du deuxième membre de (16) est l'estimateur de variance pour l'échantillon par CA de lignes fixes appliqué à la caractéristique composite  $\phi_j Y_j + \lambda(1 - \phi_j)Y_j$ , qui est la caractéristique pour les UE à téléphone fixe seulement plus une partie égale à  $\lambda$  de la caractéristique pour les UE mixtes. Le deuxième terme du deuxième membre de (16) est l'estimateur de variance pour l'échantillon de lignes de téléphone mobiles appliqué à la caractéristique composite  $\delta_j Y_j + (1 - \lambda)(1 - \delta_j)Y_j$ , qui est la caractéristique pour les UE à téléphone mobile seulement plus une partie égale à  $(1 - \lambda)$  de la caractéristique pour les UE mixtes.

Les estimateurs des matrices de covariance peuvent être construits à partir d'expressions telles que (15) et (16), ce qui facilite l'inférence statistique concernant d'autres paramètres de population d'intérêt.

## 5.4 Correction des poids d'échantillonnage

Les poids d'échantillonnage peuvent être corrigés pour tenir compte de la non-réponse ou d'un calage planifié sur des totaux de contrôle connus.

Jusqu'à présent, nous ne nous sommes pas préoccupés des divers types de données manquantes qui peuvent survenir dans une enquête par téléphone mobile. Nous nous concentrerons ici sur le calcul des corrections pour les

(12) pourraient être sujets à un biais de non-réponse différentiel qui n'est pas éliminé par la méthode classique des classes de pondération. Dans la population mixte, les utilisateurs peu fréquents du téléphone mobile pourraient être moins susceptibles de répondre s'ils sont interviewés dans l'échantillon de lignes mobiles que s'ils le sont dans l'échantillon de lignes fixes. Si ces adultes diffèrent considérablement des autres adultes dans la population mixte en ce qui concerne les caractéristiques clés étudiées dans l'enquête, l'estimateur (12), ainsi que l'estimateur (11) pour- raient présenter un biais de non-réponse.

## 5.3 Estimation de la variance

Afin de faire des inférences concernant la population globale d'après l'échantillon, nous avons besoin d'un estimateur de la variance du total estimé. Premièrement, considérons le cas de domaines non chevauchants. En travaillant dans la population d'UE, nous pouvons employer des méthodes d'estimation de la variance appropriées pour le plan de sondage. Partant de (7), le total estimé pour le domaine des unités à téléphone mobile seulement peut s'écrire

$$Y_{\text{EB}}^{\text{EB}} = \sum_{i \in \text{EB}} W_{\text{EB}}^i X_i^j$$

où

$$X_i^j = \sum_{j \in \text{EB}} \delta_j X_j^i \ell_{ij} / \sum_{j' \in \text{EB}} \ell_{ij'}. \quad (13)$$

En supposant un échantillonnage aléatoire simple, l'estimateur sans biais de la variance du total estimé est donné par

$$v(Y_{\text{EB}}^{\text{EB}}) = \sum_{i=1}^n N_h^2 \left( 1 - \frac{N_h}{n} \right) \frac{1}{s_{x_h}^2},$$

où

$$s_{x_h}^2 = \frac{1}{n} \sum_{i \in \text{EB}} \left( X_i^j - \frac{1}{n} \sum_{i' \in \text{EB}} X_{i'}^j \right)^2.$$

Si nous ignorons le facteur de correction pour population finie, ce qui serait possible dans presque toute enquête téléphonique réelle, l'estimateur de variance devient

$$v(Y_{\text{EB}}^{\text{EB}}) = \sum_{i=1}^n \frac{n_h - 1}{n_h} \sum_{i' \in \text{EB}} \left( W_{\text{EB}}^i X_i^j - \frac{1}{n_h} \sum_{i' \in \text{EB}} W_{\text{EB}}^{i'} X_{i'}^j \right)^2. \quad (14)$$

Posons maintenant que  $v(Y_{\text{EA}}^{\text{EB}})$  est un estimateur de la variance de  $Y_{\text{EA}}^{\text{EB}}$  pour l'échantillon par CA de lignes fixes. Les estimateurs de ce genre étant bien connus, nous ne les passons pas en revue ici ; voir, par exemple, Wolter et coll.



En émettant l'hypothèse d'un échantillonnage aléatoire simple sans remise dans les strates, les poids de base sont de la forme

$$(6) \quad W'_{EB} = N_h / n_h,$$

où  $h$  désigne la strate d'échantillonnage dans laquelle la  $i^e$

UE est sélectionnée,  $N_h$  est le nombre d'UE présentes dans la base de sondage qui sont comprises dans la strate  $h$ , et  $n_h$

est la taille de l'échantillon tiré dans la strate  $h$ . Habituellement, la base de sondage des lignes de téléphone mobiles

comprendrait tous les numéros de téléphone couverts par les indicateurs de central attribués par le système téléphonique

aux téléphones mobiles. L'échantillonnage aléatoire simple serait la méthode la plus courante de sélection de l'échan-

tillon pour ce genre d'indicateur. Peu d'information est disponible dans la base de sondage des lignes de téléphone

mobiles pour permettre la stratification de l'échantillon, à part l'information géographique de niveau peu détaillé

intégrée dans l'indicateur régional.

Soit  $s_{EB}^{i^e}$  l'échantillon correspondant d'UE, c'est-à-dire  $s_{EB}^{i^e} = \{j \in U_{EB} | j \text{ est liée à au moins une UE } i^e \text{ dans } s_{EB}\}$ .

Nous utiliserons cet échantillon pour estimer le total de domaine des UE qui sont liées uniquement à un téléphone

mobile,  $Y_{EB}$ . Partant de (1) et (2), il est facile de voir que l'estimateur sans biais du total de domaine est donné par

$$Y_{EB} = \sum_{i^e \in U_{EB}} W'_{EB} \left\{ \sum_{j \in U_{EB}} \delta_j Y_j \ell_{ij} / \sum_{j \in U_{EB}} \ell_{ij} \right\}$$

$$(7) \quad = \sum_{i^e \in U_{EB}} \delta_{i^e} Y_{i^e} W'_{EB},$$

où les poids d'échantillonnage au niveau des UE sont définis par

$$(8) \quad W'_{EB} = \sum_{i^e \in U_{EB}} W'_{EB} \ell_{ij} / \sum_{j \in U_{EB}} \ell_{ij} \text{ pour } j \in s_{EB}.$$

De nouveau, voir Lavallée (2007) pour l'expression de ces poids dans le contexte de l'échantillonnage indirect.

Avant de quitter le domaine  $b$ , nous observons en passant qu'il est possible de sous-échantillonner les UE et de recueillir l'information d'enquête uniquement pour le sous-

échantillon au lieu de recenser toutes les UE liées aux UR de l'échantillon. Si le statisticien choisissait une certaine forme

d'échantillon ou le coût, un facteur de pondération supplémentaire apparaîtrait dans les poids donnés par (8). Ce

genre de sous-échantillonnage est dénommé échantillonnage indirect à deux degrés dans Lavallée (2007, section 5.1).

Tournons-nous vers le domaine  $A$ . Soit  $s_{EA}^{i^e}$  un échantillon par CA standard de lignes de téléphone fixes, soit  $s_{EA}^{i^e}$

l'échantillon connexe d'UE, c'est-à-dire  $s_{EA}^{i^e} = \{j \in U_{EA} | j \text{ est liée à au moins une UE } i^e \text{ dans } s_{EA}^{i^e}\}$ , et soit

$$(9) \quad Y_{EA} = \sum_{i^e \in U_{EA}} W'_{EA} Y_{i^e}$$

l'estimateur sans biais standard du total de population. Par souci de concision, nous ne dériverons pas les poids d'échantillonnage standard ici ; pour plus de renseignements au sujet

de ces poids, consulter Wolter et coll. (2008).

De (7) et (9) il découle que l'estimateur sans biais du total de population des UE est donné par

$$(10) \quad Y_{ET} = Y_{EA} + Y_{EB}$$

## 5.2 Cas des domaines chevauchants

Procédons maintenant à l'estimation en partant de la décomposition (4). Cela signifie que, dans l'échantillon de

lignes de téléphone mobiles, nous interviewerons non seulement la population ne possédant qu'un téléphone mobile,

mais aussi la population mixte (c'est-à-dire les personnes qui utilisent à la fois un téléphone fixe et un téléphone

mobile). L'estimateur du total de population d'intérêt est maintenant de la forme

$$(11) \quad Y_{ET} = Y_{EA} + Y_{EAB} + Y_{EB},$$

où

$$Y_{EA} = \sum_{i^e \in U_{EA}} W'_{EA} \phi_{i^e} Y_{i^e}$$

est l'estimateur pour le domaine des unités à téléphone fixe seulement dérivées de l'échantillon de lignes de téléphone fixes,  $Y_{EB}$  est défini en (7) et est l'estimateur du domaine

des unités à téléphone mobile seulement dérivées de l'échantillon de lignes de téléphone mobiles, et  $Y_{EAB}$  est un estimateur du domaine mixte obtenu à partir des deux échan-

tillons. L'estimateur pour le domaine mixte est

$$Y_{EAB} = \lambda \sum_{i^e \in U_{EA}} W'_{EA} (1 - \phi_{i^e}) Y_{i^e}$$

$$(12) \quad + (1 - \lambda) \sum_{j \in U_{EB}} W'_{EB} (1 - \delta_j) Y_j.$$

Les poids nécessaires pour appuyer l'estimateur (11) sont  $\{W'_{EA}\}$  et  $\{W'_{EB}\}$ .

Voit Hartley (1962) pour une discussion du paramètre de mélange  $\lambda$  dans une enquête à base de sondage double, en

se concentrant sur des considérations de variabilité d'échantillonnage. Si nous nous préoccuons maintenant du biais,

Brick et coll. (2006) signalent que la propension à répondre à une enquête par téléphone mobile pourrait être liée positivement à la fréquence d'utilisation du téléphone

mobile. Donc, les deux éléments du deuxième membre de



pour désigner cette sous-population ayant le téléphone. Sub-séquentement, à la section 5.4, nous discutons brièvement de la couverture de la population sans téléphone.

Pour les UE comprises dans  $U^E$ , définissons les variables indicatrices

$$\delta_j = 1, \text{ si aucune des UR liées à } j \text{ n'a accès à un service de téléphone fixe, tandis qu'au moins une des UR a habituellement accès à un service de téléphone mobile ;}$$

$$= 0, \text{ autrement.}$$

$\phi_j = 1$ , si aucune des UR liées à  $j$  n'a habituellement accès à un service de téléphone mobile, tandis qu'au moins l'une des UR a accès à un service de téléphone fixe ;

$= 0$ , autrement.

La variable  $\delta$  est indicatrice de la situation de possession d'un téléphone mobile seulement et la variable  $\phi$  est indicatrice de la situation de possession d'un téléphone fixe seulement.

Alors, le total de population d'intérêt peut être décomposé en

$$Y_{ET} = Y_{EA} + Y_{EB}, \quad (3)$$

où

$$Y_{Eb} = \sum_{j \in U_{Et}} \delta_j Y_j$$

est le total pour le domaine des unités à téléphone mobile seulement, et

$$Y_{Ea} = \sum_{j \in U_{Et}} (1 - \delta_j) Y_j$$

est le total du complètement de ce domaine, qui comprend les UE liées exclusivement à des lignes fixes et les UE mixtes, qui sont liées à la fois à des lignes de téléphone fixes et mobiles. Le total des UE peut également s'écrire

$$Y_{ET} = Y_{Ea} + Y_{Eab} + Y_{Eb}, \quad (4)$$

où

$$Y_{Ea} = \sum_{j \in U_{Et}} \phi_j Y_j$$

est le total pour la population d'unités à téléphone fixe seulement, et

$$Y_{Eab} = \sum_{j \in U_{Et}} (1 - \delta_j) (1 - \phi_j) Y_j$$

est le total pour la population mixte ayant accès à une combinaison de téléphones fixes et de téléphones mobiles. Enfin, le total de population peut s'écrire

### 5.1 Cas des domaines non chevauchants

À la présente section, nous utilisons un échantillon de lignes mobiles en vue de produire une estimation pour la population ayant un téléphone mobile seulement  $U_{Eb}$  et un échantillon de lignes fixes pour la production d'estimations pour la population complète de lignes de téléphone fixes  $U_{Ea}$ . Nous observons qu'il est impossible de sélectionner directement un échantillon d'unités ayant une ligne de téléphone mobile seulement, parce que l'information sur la situation d'unité à téléphone mobile seulement n'est pas disponible dans la base de sondage, mais est plutôt recueillie au moyen du questionnaire de présélection de l'enquête. Pour appliquer ce plan, on éliminerait par présélection les répondants ayant un téléphone mobile qui se classent eux-mêmes dans le domaine mixte et on terminerait l'interview, en poursuivant uniquement avec les répondants possédant un téléphone mobile seulement.

Soit  $s_{Eb}$  un échantillon probabiliste d'UE (lignes de téléphone mobiles) sélectionné dans la population  $U_{Eb}$ , et soit  $\{W'_{Eb}\}$  l'ensemble de poids d'échantillonnage de base, tels que

$$X_{Eb} = \sum_{i \in s_{Eb}} W'_i X_i$$

est un estimateur sans biais du total de population  $X_{Eb}$ , où  $X_i$  est une caractéristique de la  $i^e$  unité de la population.

où

$$Y_{ET} = Y_{Ea} + Y_{Eb}, \quad (5)$$

$$Y_{Eb} = \sum_{j \in U_{Et}} (1 - \phi_j) Y_j$$

#### 4. Dualité entre les populations d'UE et d'UE

Afin d'entreprendre l'élaboration d'une méthode d'estimation sans biais pour les enquêtes par téléphone mobile, nous établissons qu'une dualité existe entre la population d'UE, ou téléphones mobiles, (donc désignées par  $U^{EB}$ ) et la population d'UE qui sont liées aux téléphones mobiles (désignées par  $U^{EB}$ ). L'objectif d'une enquête par téléphone mobile est de faire des inférences concernant  $U^{EB}$ , mais nous verrons bientôt que cet objectif équivaut à faire certaines inférences concernant  $U^{EB}$  (dans cette notation, la première lettre de l'indice supérieur désigne le type d'unité, tandis que la lettre  $B$  désigne la base de sondage des lignes de téléphone mobiles. Plus loin, nous utiliserons l'indice  $A$  pour désigner la base de sondage des lignes de téléphone fixes).

Dans le domaine des UE, un total de population d'intérêt est donné par

$$Y^{EB} = \sum_{i \in I^{EB}} Y_i$$

où, dans le deuxième membre, la variable  $Y$  est un item du questionnaire ou une autre variable, recodée ou dérivée, attachée aux unités de la population  $U^{EB}$ . De même, dans le domaine des UE, un total de population est défini par

$$X^{EB} = \sum_{i \in I^{EB}} X_i$$

où, dans le deuxième membre, la variable  $X$  représente toute caractéristique fixe attachée aux unités de la population  $U^{EB}$ .

Alors que l'analyste des données de l'enquête s'intéresse au total pour la population d'UE (et à d'autres paramètres de cette population), nous pouvons obtenir un paramètre correspondant dans le domaine des UE en écrivant

$$Y^{EB} = \sum_{i \in I^{EB}} Y_i = \sum_{i \in I^{EB}} \sum_{j \in I^{EB}} \frac{Y_j}{X_j} X_j = \sum_{j \in I^{EB}} X_j \cdot Y_j \quad (1)$$

où la variable  $X$  est maintenant définie spécifiquement par

$$X_j = \sum_{i \in I^{EB}} \frac{Y_j}{X_j} \quad (2)$$

L'expression (1) montre la correspondance entre l'estimation dans le domaine des UE et l'estimation dans le domaine des UE, avec  $X_j$  définie comme en (2), est équivalent au total d'intérêt  $Y^{EB}$ , d'où le problème de l'estimation de  $Y^{EB}$  est équivalent au problème de l'estimation de  $X^{EB}$ .

Nous notons que (2) se présente essentiellement sous la même forme dans la théorie de l'échantillonnage indirect. Voir Lavallée (2007), Théorème 4.1. En échantillonnage

#### 5. Estimation

indirect, les UE sont liées à des grappes naturellement définies d'UE; si une UE donnée est sélectionnée dans l'échantillon, les données d'enquête sont recueillies pour toutes les UE comprises dans les grappes qui y sont liées. Ici, l'analogue est que les grappes sont définies par les UR qui répondent à la tentative d'interview par téléphone mobile et les données d'enquête sont recueillies auprès du répondant pour toutes les UE auxquelles il est lié. Ici, la situation est telle que la grappe est définie par la paire UE-UR. À cet égard, il se pose un problème d'identifiabilité qui n'existe généralement pas en échantillonnage indirect et nous examinons cette question plus en détail à la section 5.5.

Dans (2), nous attribuons effectivement une part égale de la variable  $X_j$  à chaque UE  $i$  à laquelle la variable est liée. Nous pourrions aussi aboutir au même résultat en répartissant  $X_j$  entre les UE auxquelles elle est liée, proportionnellement à une autre mesure connue de l'intensité de la relation entre  $j$  et  $i$ . Nous pourrions certes concevoir une répartition optimale de  $X_j$  entre les UE qui sont liées, comme dans Deville et Lavallée (2006), mais ce genre de répartition peut être difficile à exécuter ou pourrait ne pas avoir beaucoup de poids dans des conditions pratiques à grande échelle.

Comme nous l'avons mentionné dans l'introduction, certaines UE seront liées exclusivement à des lignes de téléphone mobiles, certaines seront liées exclusivement à des lignes de téléphone fixes et certaines seront liées à la fois à des lignes de téléphone fixes et mobiles. Les UE sans téléphone, si tant est qu'il y en ait, ne seront liées ni à des lignes de téléphone mobiles ni à des lignes fixes. Pour établir la notation dans ce contexte, posons que  $U^E$  est la population globale d'UE d'intérêt et que  $U^E$  est la population globale d'UE. Soit  $U^{EA}$  les éléments de  $U^E$  qui sont liés à des lignes fixes, soit  $U^{EB}$  les éléments qui sont liés à des lignes mobiles, soit  $U^{Ea}$  les éléments qui sont liés uniquement à des lignes fixes, soit  $U^{Eb}$  les éléments qui sont liés uniquement à des lignes mobiles, soit  $U^{Eab}$  les éléments qui sont liés à la fois à des lignes fixes et à des lignes mobiles, soit  $U^{Ea} = U^{Ea} \cup U^{Eb}$ , et  $U^{Eb} = U^{Ea} \cup U^{Eb}$ , ou  $U^{Ea}$  et  $U^{Eb}$  sont des ensembles disjoints. En outre, soit  $U^{EA}$  la population de lignes de téléphone fixes, telles que  $U^E = U^{EA} \cup U^{EB}$ . Les lignes fixes et les lignes mobiles relèvent des sous-ensembles disjoints de la population globale d'UE. Aux sections 5.1 et 5.2 qui suivent, nous discutons de l'estimation sans biais pour la sous-population, disons  $U^{ET} = U^{EA} \cup U^{EB}$ , qui est liée à au moins un téléphone et n'importe quelle sorte. Nous utilisons l'indice supérieur  $T$



Nous disons qu'un adulte donné a un accès habituel à un NMPA si, et uniquement si, cette personne possède l'usage

- régulier ;
- important ;
- continu de la ligne de téléphone mobile.

Chaque NMPA possède un ou plusieurs utilisateurs adultes réguliers et chaque utilisateur possède l'accès habituel à un ou à plusieurs téléphones mobiles. Dans de nombreux cas, il existe une relation un à un unique entre la ligne de téléphonie mobile et l'utilisateur adulte. Dans certains cas, par contre, il existe une relation un à plusieurs entre la ligne de téléphone mobile et ses utilisateurs.

Nous traitons une UE donnée et une UR donnée comme étant liées si, et uniquement si, l'UE est un NMPA et que l'UR a habituellement accès à l'UE. Une enquête par téléphone mobile doit tenir compte des liens qui existent entre la population d'UE et la population d'UR et reconnaître ces liens.

3.2 Lien entre l'UE et l'UR

Une UE donnée est liée à une ou à plusieurs UR par les relations naturelles qui existent dans la société, telles que celles créées par la famille ou le lieu de résidence. Par exemple, un répondant adulte peut répondre à l'interview de l'enquête au nom de son ménage, de sa famille ou d'une unité de consommation. Il peut répondre en son propre nom, pour un enfant à charge de moins de 18 ans ou pour ses parents ou ses frères et sœurs.

Toute enquête nécessite un protocole de réponse qui précise quels répondants adultes doivent répondre pour quelles UE. Le protocole est choisi par le méthodologiste de l'enquête en tenant compte de la faisabilité, du coût et de l'exactitude de la déclaration. C'est ce protocole qui établit les liens entre les UE et les UR.

3.3 Lien entre l'UE et l'UE

Les liens qui précèdent entre les UR et les UE, ainsi qu'entre les UE et les UR déterminent les liens entre les UE et les UE. Nous disons qu'une UE donnée est liée à une UE donnée si, et uniquement si, l'UE est liée à au moins une UR qui, à son tour, est liée à l'UE.

La notation qui suit sera utile dans l'exposé des sections suivantes. Soit  $j$  une UE donnée dans la population d'intérêt et soit  $i$  une UE donnée dans la population. Alors, définissons les variables indicatrices ou variables de lien

$$l_{ij} = \begin{cases} 1, & \text{si la } j^{\text{e}} \text{ UE est liée à la } i^{\text{e}} \text{ UE} \\ 0, & \text{autrement.} \end{cases}$$

Unités d'échantillonnage (UE) – Lignes téléphoniques	Unités répondantes (UR) – Adultes	Unités d'estimation (UE) – Ménages, adultes ou enfants
--	-----------------------------------	--

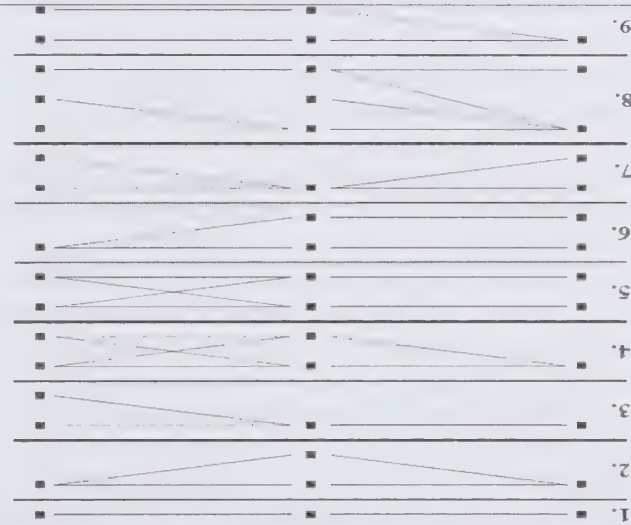


Figure 1 Exemples de réseaux dans une enquête par téléphone mobile

3. Liens entre les unités dans le réseau

Un lien est une relation saillante entre deux nœuds du réseau. Dans le contexte de la figure 1, les liens sont représentés par les segments de droite qui unissent les différents nœuds. Afin de fournir un fondement pour l'estimation d'après les données d'enquête, nous devons explorer les liens entre i) les UR et les UE, ii) les UE et les UR, et iii) les UE et les UE.

3.1 Lien entre l'UR et l'UE

Deux concepts sont essentiels à la création d'un lien entre une UR et une UE, à savoir les concepts a) de *numéro de téléphone mobile personnel actif* (NMPA) et b) d'*accès habituel* à la ligne mobile.

Un NMPA correspond à une ligne téléphonique qui est en service au moment de l'enquête par téléphone mobile et dont les appels parviennent à un adulte admissible qui utilise le téléphone mobile, au moins partiellement, pour des raisons personnelles. Autrement dit, un NMPA doit satisfaire à trois critères :

- il est en service ;
- il permet d'entrer en communication avec un répondant adulte admissible ;
- il n'est pas utilisé exclusivement pour affaires.



équivalaient à un échantillonnage en réseau. À la section 3, nous présentons divers concepts dont nous aurons besoin durant la discussion de l'estimation d'après les données de sondage, entre autres la notion d'un *lien* (ou arête) entre les *nœuds* (ou sommets) du réseau. À la section 4, nous décrivons la dualité qui existe entre les populations correspondantes aux différents types de nœuds. Notre approche rappellera à certains lecteurs les méthodes d'échantillonnage indirectes de Lavallée (2007). Le cœur de l'article est la section 5, où nous présentons les estimateurs sans biais des totaux de population pour les enquêtes par téléphone mobile et pour les plans d'enquête téléphonique à base de sondage double correspondants. À la section 6, à titre d'exemple, nous illustrons les incidences des nouvelles méthodes d'estimation dans le cas d'une enquête téléphonique existante sur la couverture de la vaccination des jeunes enfants et des adolescents. Nous concluons, à la section 7, par un bref résumé.

2. Réseaux d'unités et le protocole de réponse

En général, au moins trois types d'unités existent dans le contexte d'une enquête par téléphone mobile, à savoir :

- les unités d'échantillonnage (UE) ;
- les unités répondantes (UR) ;
- les unités d'estimation (UE).

L'UE est l'unité d'échantillonnage de l'enquête. En pratique, les numéros de téléphone peuvent être échantillonnés directement d'après des listes de numéros de téléphone mobile ou l'être par étape, les indicatifs de central ou les banques de numéros servant d'unités primaires d'échantillonnage, et les numéros proprement dits étant sélectionnés en une ou plusieurs étapes de sous-échantillonnage dans les unités primaires. Pour simplifier la discussion, dans le présent article, nous considérons le numéro de téléphone proprement dit comme étant l'UE.

La cible réelle de l'interview de l'enquête, c'est-à-dire l'unité d'analyse, est ce que nous appelons l'unité d'estimation (UE). Certaines enquêtes sont axées sur la collecte et l'analyse de données sur les ménages ou sur les familles, le ménage ou la famille étant alors l'UE. D'autres visent à recueillir des données au niveau de la personne, auquel cas les personnes admissibles peuvent être des enfants de moins

de 18 ans, des adultes de 18 ans et plus, ou un segment démographique particulier de la population, tel que les femmes hispaniques de 0 à 34 ans. D'autres enquêtes encore sont axées sur les données au niveau du ménage ainsi qu'au niveau de la personne, et elles comportent dans ce cas au moins deux types d'UE et deux niveaux d'analyse.

Dans les enquêtes téléphoniques, l'adulte est l'unité répondante ou UR. L'UE peut ou non avoir la capacité de fournir directement les réponses à son sujet. Dans la négative, une UR répond en son nom. Si l'UE est un adulte, celui-ci, voire même un autre adulte peut servir d'UR correspondante. Si l'UE est un ménage, une famille, une unité de consommation ou un enfant, un ou plusieurs adultes peuvent servir d'UR correspondante. Le protocole de réponse, qui est spécifié par le méthodologiste de l'enquête, détermine effectivement quelles UR ont le droit de répondre pour quelles UE. Dans une enquête téléphonique habituelle, pour chaque UE sélectionnée dans l'échantillon, on communique par téléphone et on interviewe un répondant adulte (ou UR).

Dans une enquête par téléphone mobile, les liens entre les UE, les UR et les UE peuvent être différents. La figure 1 présente neuf réseaux qui illustrent certains types de liens possibles. Dans le premier réseau, une UE est liée à une UR, qui à son tour répond pour une UE. Cet arrangement pourrait avoir lieu si un adulte utilise une ligne téléphonique et que cet adulte à son tour fait la déclaration pour le deuxième réseau, une UE est liée à deux UR, qui peuvent répondre chacune pour l'UE. Cet arrangement s'observerait, par exemple, si deux adultes partageaient la même ligne téléphonique et que chacun aurait le droit, en vertu du protocole de l'enquête, de répondre au nom du ménage. Le cinquième réseau pourrait avoir lieu si deux adultes possédaient chacun leur propre ligne téléphonique non partagée avec l'autre, mais que chaque adulte de la paire avait le droit, en vertu du protocole de l'enquête, de répondre pour chacun des deux enfants.

Des réseaux plus compliqués sont possibles et doivent certainement exister de par le monde. Par exemple, le huitième réseau représente un arrangement où trois adultes partagent deux lignes téléphoniques. La première des lignes est partagée par les trois adultes, tandis que la deuxième est utilisée uniquement par le troisième adulte. Le premier des adultes a le droit, en vertu du protocole de l'enquête, de répondre pour deux UE, telles que ses enfants biologiques ; le deuxième n'a le droit de répondre pour aucune UE, et le troisième a le droit de répondre uniquement pour une troisième UE pour laquelle les deux premiers adultes ne peuvent pas faire de déclaration.

par ligne fixe ou par ligne mobile sont considérées comme n'ayant pas de téléphone. Selon les données courantes, bien que personne n'en soit certain, de 20 % à 30 % environ des adultes font partie du domaine *b*, de 5 % à 10 %, du domaine *c*, et les autres sont répartis entre les domaines *a* et *ab*.

Ce que nous ont appris jusqu'à présent nos enquêtes par téléphone mobile et celles réalisées par d'autres est que la collecte des données est relativement coûteuse, le nombre moyen d'heures d'intervieweur par cas achevé étant de l'ordre de trois fois la moyenne observée pour les enquêtes conventionnelles par CA. Le coût plus élevé est dû, en partie, à l'exigence légale (aux États-Unis, la *Telephone Consumer Protection Act*) de composer manuellement les numéros de téléphone mobile sélectionnés. Les taux de réponse sont un peu plus faibles que ceux obtenus dans les enquêtes par CA. La durée de l'interview peut poser problème, certains répondants étant moins disposés à répondre à une longue interview par téléphone mobile que par téléphone fixe. En outre, des questions de protection de la vie privée peuvent constituer une contrainte dans le cas de l'interview par téléphone mobile, si la personne ne se trouve pas dans un lieu privé au moment de l'interview. La propension des utilisateurs de téléphone mobile à répondre peut varier de manière monotone avec leur niveau d'utilisation du téléphone mobile, les gros utilisateurs étant plus disposés à répondre que les utilisateurs légers ou occasionnels. La plupart des ruptures ont lieu durant les premières secondes de la tentative d'interview. Comme les enquêtes par téléphone mobile sont relativement nouvelles, les appels des intervieweurs sont inhabituels et ceux-ci n'ont que quelques secondes pour vendre l'enquête. Par ailleurs, nous constatons que de nombreux répondants rejoignent sur téléphone mobile sont relativement prêts à coopérer une fois que l'on a réussi à capter leur attention en leur lisant l'introduction à l'enquête.

Étant donné toutes ces circonstances, nous considérons à l'heure actuelle l'échantillon de lignes mobiles comme un échantillon complètement relativement petit, l'échantillon principal continuant d'être un échantillon par CA plus grand de lignes fixes. L'échantillon de lignes mobiles est destiné à paraître la couverture de la population d'intérêt. Dans l'avenir, à mesure que le contexte des enquêtes par téléphone mobile évoluera vers la maturité et que les coûts baisseront, il sera peut-être possible de passer à une approche plus équilibrée où les échantillons de lignes fixes et de lignes mobiles seront de tailles similaires, voire même une situation où l'échantillon de lignes mobiles commun-cera à devenir dominant et où l'échantillon de lignes fixes sera utilisé comme complètement pour achever la couverture.

À la section 2, nous décrivons les réseaux d'unités d'échantillonnage, d'unités répondantes et d'unités d'estimation, et montrons que les enquêtes par téléphone mobile

figurant dans les deux bases de sondage sont non chevauchantes, tandis que les personnes et les ménages correspondants qui peuvent être les sujets de l'enquête sont partiellement chevauchants.

Une théorie rigoureuse de l'estimation pour ce genre de plan de sondage téléphonique fait défaut, quoique certaines descriptions initiales de la pondération aient été avancées par Brick, Dipko, Presser, Tucker et Yuan (2006), Brick, Edwards et Lee (2007), et Frankel, Battaglia, Link et Mokdad (2007). Dans le présent article, nous présentons une théorie générale de l'estimation sans biais des totaux de population dans le contexte de plans d'enquête téléphonique à base de sondage double et calculons les poids de sondage correspondants. Nous montrons que l'information doit être recueillie durant l'enquête proprement dite pour permettre le calcul des poids d'échantillonnage.

Après de fixer les idées, posons que *A* désigne la partie de la population globale d'intérêt accessible par la voie de la base de sondage des lignes de téléphone fixes, *B* désigne la partie accessible par la voie de la base de sondage des lignes de téléphone mobiles et *C* désigne la partie qui n'est accessible par la voie d'aucune des deux bases de sondage (la *population sans téléphone* et d'autres composantes relativement petites de la population totale). Soit *a* la sous-population dans *A* qui n'est pas accessible par la voie des lignes de téléphone mobiles (la *population à téléphone fixe seulement*), soit *b* la sous-population dans *B* non accessible par la voie des lignes téléphoniques fixes (la *population à téléphone mobile seulement*), et soit *ab* la sous-population accessible à la fois par ligne fixe et par ligne mobile (la *population à accès mixte*). Nous peaufinerons cette notation dans les sections qui suivent.

Savoir si une unité de la population d'intérêt est accessible par des lignes fixes ou des lignes mobiles est en soi une question complexe. Tout au long du présent article, quand nous disons qu'une unité est accessible par ligne fixe, nous entendons à la fois qu'il existe un accès à une ou à plusieurs lignes fixes (habituellement des lignes terrestres résidentielles uniquement) et qu'une personne répondrait effectivement à l'appel sur le téléphone fixe s'il sonnait pour des communications vocales. De nos jours, de nombreux adultes gardent une ligne téléphonique fixe strictement pour les communications informatiques et utilisent un téléphone mobile pour toutes les communications vocales. En vertu de notre définition, ces adultes ne sont pas considérés comme ayant accès à une ligne fixe et sont au contraire considérés comme faisant partie de la population à téléphone mobile accessible par ligne mobile, nous entendons qu'il existe à la fois l'accès physique à un téléphone mobile et l'intention de répondre à ce téléphone mobile s'il sonne. Toutes les autres unités de la population d'intérêt qui ne sont pas accessibles



# Fondements statistiques des enquêtes par téléphone mobile

Kirk M. Wolter, Phil Smith et Stephen J. Blumberg<sup>1</sup>

## Résumé

Aux États-Unis, la taille de la population dotée d'un téléphone mobile augmente rapidement ces dernières années et, par conséquent, les chercheurs ont commencé à expérimenter l'échantillonnage et l'interview des abonnés à la téléphonie mobile. Nous discutons des problèmes statistiques que posent les étapes de l'établissement du plan d'échantillonnage et de l'estimation des études par téléphone mobile. Les travaux sont exposés principalement dans le contexte d'une enquête à deux bases de sondage non chevauchantes dans laquelle une base et un échantillon sont employés pour la population possédant un téléphone fixe et une deuxième base de sondage et un deuxième échantillon, pour la population possédant uniquement un téléphone mobile. Nous discutons également des aspects supplémentaires dont il faut tenir compte dans le cas d'une enquête à deux bases de sondage chevauchantes (où la base de sondage et l'échantillon pour la téléphonie mobile comprennent certains membres de la population dotée d'un téléphone fixe). Pour illustrer les méthodes, nous utilisons le plan de sondage de la National Immunization Survey (NIS) conçu pour surveiller les taux de vaccination chez les enfants de 19 à 35 mois et les adolescents de 13 à 17 ans. La NIS est une enquête téléphonique nationale, suivie d'une vérification des dossiers des fournisseurs de service, réalisée par les Centers for Disease Control and Prevention.

Mots clés : Étude par téléphone mobile ; composition aléatoire ; enquête à base de sondage double ; échantillonnage en réseau ; échantillonnage indirect ; règles de lien ; pondération des données d'enquête ; National Immunization Survey.

## 1. Introduction

Aux États-Unis, le nombre de personnes possédant un téléphone mobile a augmenté rapidement ces dernières années et le pourcentage d'adultes vivant dans un ménage équipé de téléphones mobiles devrait rapidement surpasser celui des adultes vivant dans un ménage muni de téléphones fixes (CTIA 2008 ; Blumberg et Luke 2008 ; Arthur 2007 ; Ehlen et Ehlen 2007). Par conséquent, les spécialistes de la recherche sur les sondages ont commencé à expérimenter des méthodes d'échantillonnage et d'interview des abonnés à la téléphonie mobile (Lavrakas, Shuttles, Steeh et Fienberg 2007). Le présent article porte sur les problèmes de conception et d'estimation statistique que posent les enquêtes par téléphone mobile. Y sont mises en relief des solutions rigoureuses sur le plan théorique, mais pratiques aux nouveaux problèmes que doivent résoudre aujourd'hui les spécialistes de la recherche sur les sondages dans le cadre des enquêtes par téléphone mobile.

L'enquête par téléphone mobile entraîne un changement de paradigme et pose de nouveaux défis. Dans l'esprit de la plupart des gens, un téléphone mobile est un appareil personnel et non un appareil ménager. Certaines personnes partagent un téléphone mobile, dont 10 % à 20 % des adultes munis seulement d'un téléphone mobile (Carley-Baxter, Peytchev et Lynberg 2008), mais bon nombre ne le font pas et l'on ne peut donc pas supposer que tous les membres d'un ménage peuvent être rejoints au moyen de la même ligne téléphonique mobile. En outre, certains membres d'un ménage peuvent être rejoints au moyen de plus d'une ligne de téléphone mobile. Enfin, certains membres du ménage peuvent être rejoints uniquement au moyen d'une ligne de téléphone mobile, tandis que d'autres peuvent l'être au moyen d'un téléphone mobile et d'un téléphone fixe. Donc, dans les enquêtes par téléphone mobile, le ménage ne représente plus forcément la même organisation unitaire que dans le cas des enquêtes par téléphone conventionnelle.

Afin de lutter contre le risque croissant de biais (dû au sous-dénombrement) dans les enquêtes téléphoniques, on peut considérer des plans d'enquête téléphonique à base de sondage double qui comprennent un échantillon sélectionné par CA de lignes de téléphone fixes et un échantillon de lignes de téléphone mobiles. Les numéros de téléphone

<sup>1</sup> Kirk M. Wolter, NORC et University of Chicago. Courriel : wolter-kirk@norc.org ; Phil Smith, National Center for Immunization and Respiratory Diseases ; Stephen J. Blumberg, National Center for Health Statistics.



## Conclusion

Nous avons étudié l'estimation de la variance totale des estimateurs des paramètres du modèle sous l'hypothèse d'un modèle de superpopulation. Notre approche mène directement à un estimateur de variance par linéarisation qui, comme nous le montrons, donne de bons résultats dans un cadre conditionnel quand les poids de calage sont utilisés pour l'estimation. Nous sommes en train d'étudier des extensions de notre méthode à l'estimation de la variance totale sous imputation pour la non-réponse partielle et sous intégration de deux enquêtes indépendantes.

## Remerciements

Nous remercions deux examinateurs de leurs suggestions et commentaires constructifs. Les travaux de J.N.K. Rao ont été financés en partie par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada.

## Bibliographie

- Demati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête (avec discussion). *Techniques d'enquête*, 30, 17-37.
- Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.
- Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*. New York : John Wiley & Sons, Inc.
- McCullagh, P., et Nelder, J.A. (1989). *Generalized Linear Models*, 2<sup>ème</sup> Ed. Chapman & Hall, Londres.
- Molina, E.A., Smith, T.M.F., et Sugden, R.A. (2001). Modeling overdispersion for complex survey data. *Revue Internationale de Statistique*, 69, 373-384.
- Royall, R.M., et Cumberland, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.
- Rubin-Bleuer, S., et Schiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, 33, 2789-2810.
- Sämdal, C.-E., Swensson, B., et Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- Skinner, C.J., Holt, D., et Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York : John Wiley & Sons, Inc.
- Valliant, R., Dorfman, A.H., et Royall, R.M. (2000). *Finite population sampling and inference: A prediction approach*. New York : John Wiley & Sons, Inc.
- Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.
- Binder, D. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases : une approche de type « recette ». *Techniques d'enquête*, 22, 17-22.

première,  $est(I) = g_s(\theta_1)$ , donne lieu à une sous-estimation très importante, comme prévu.

Nous avons également examiné les propriétés conditionnelles des trois estimateurs de variance de la même façon qu'à la section 2.2. Nous avons classé les 40 000 échantillons par ordre croissant de taille,  $n_1$ , dans la classe 1, puis les avons groupés en 20 groupes, chacun de taille 2 000, de manière que le premier groupe,  $G_1$ , contienne les 2 000 échantillons ayant les plus petites valeurs  $n_1$ , que le deuxième,  $G_2$ , contienne les 2 000 échantillons ayant les plus petites valeurs  $n_1$ , et ainsi de suite, pour obtenir 20 groupes,  $G_1, \dots, G_{20}$ .

Nous avons calculé l'EQM conditionnelle de  $\theta_1$  et le biais relatif conditionnel (BRC) connexe des estimateurs de variance  $\theta_{DR}^{cus}$ ,  $\theta_{cus}$  et  $\theta_{mix}$  basés sur les valeurs moyennes de  $\theta_{DR}^{cus}$ ,  $\theta_{cus}$  et  $\theta_{mix}$  dans chaque groupe ; voir la figure 5.

L'examen de cette figure montre que le BRC de  $\theta_{cus}$  varie de 20 % à -20 % lorsque l'on passe d'un groupe à l'autre, tandis que  $\theta_{DR}^{cus}$  ne présente pas ce genre de tendance et son BRC est inférieur à 5 % en valeur absolue, sauf pour deux groupes. En outre, le BRC de  $\theta_{mix}$  présente une tendance, mais moins prononcée que celle du BRC de  $\theta_{cus}$ . La figure 6 donne les taux de couverture conditionnels (TCC) des intervalles de la théorie normale basés sur  $\theta_{DR}^{cus}$  et  $\theta_{mix}$  pour le niveau nominal de 95 %. La figure 6 montre que  $\theta_{cus}$  présente une tendance selon le groupe, le TCC variant de 97 % à 92 %, tandis que le TCC associé à  $\theta_{DR}^{cus}$  est proche du niveau nominal dans les divers groupes. En outre, le TCC associé à  $\theta_{mix}$  est légèrement supérieur à celui associé à  $\theta_{DR}^{cus}$  pour la première moitié des groupes et légèrement inférieur, pour les autres.



Figure 5 Biais relatif conditionnel des estimateurs de variance : régression logistique

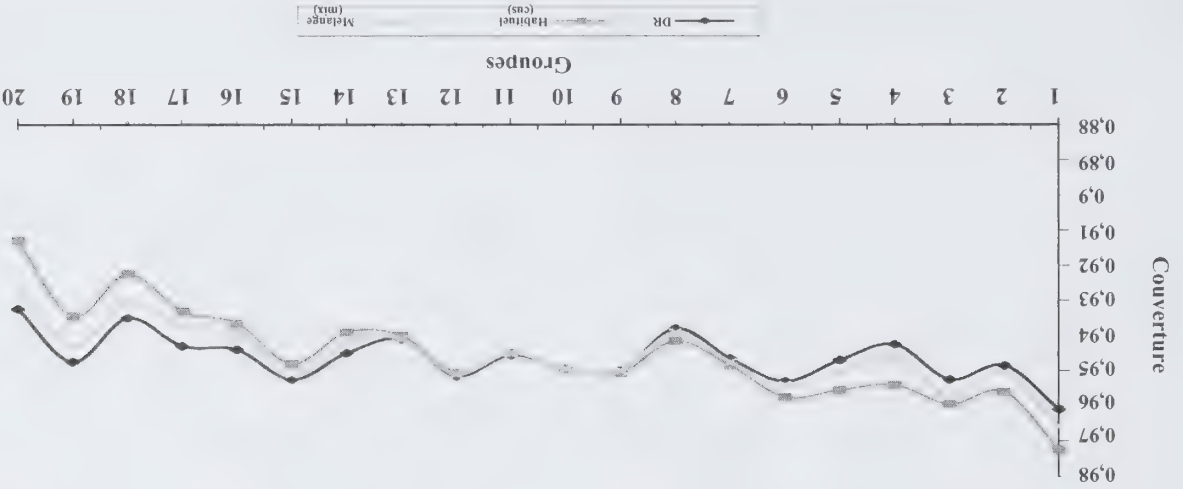


Figure 6 Taux de couverture conditionnel des intervalles de confiance de la théorie normale pour le niveau nominal de 95 % : régression logistique

au cas vectoriel  $U = \sum U^k d^k(s) = \sum b_T^k d^k(s)$ , où  $b^k = U^k h^k$  est un  $p$ -vecteur et  $U^k$  est une matrice  $p \times (p+1)$  dont les lignes sont  $u_T^{jk}$ ,  $j = 1, \dots, p$ . Dans ce cas, l'estimateur de variance SYG (2.4) devient

$$\text{est}(I) = g^{\text{SYG}}(U) = \sum \sum_{k < l} d_{kl}^k(s) \frac{\pi_k \pi_l}{(\pi_k \pi_l - \pi_{kl})} (b^k - b^l)(b^k - b^l)^T. \quad (3.2)$$

De même, l'estimateur de variance HT (2.5) devient

$$\text{est}(I) = g^{\text{HT}}(U) = \sum \sum d_{kl}^k(s) \frac{(\pi_{kl} - \pi_k \pi_l)}{(\pi_k \pi_l - \pi_{kl})} b_{kl}^k b_{kl}^T. \quad (3.3)$$

Si nous passons à la composante  $II$  de la variance totale de  $U$ , (2.6) devient

$$\text{est}(II) = \sum \sum d_{kl}^k(s) U^k \text{cov}_m(h^k, h^l) U^{lT}. \quad (3.4)$$

La variance totale de  $U$  est estimée par la somme de

(3.2) et (3.4) pour les plans à taille d'échantillon fixe ou par la somme de (3.3) et (3.4) pour les plans arbitraires.

Un estimateur de variance par linéarisation de la variance totale de  $\tilde{\theta}$  s'obtient à partir de l'estimateur de variance

linéarisée  $Z_k = \partial f(A_b) / \partial \theta_k |_{A_b = A_j}$ . Si nous suivons la méthode de dérivation implicite de Dermati et Rao (2004),

$Z_k$  se réduit à

$$Z_k = [J(\tilde{\theta})]^{-1} g_k(d(s)) (-b_T^l t_k, I^p),$$

avec

$$b_l^k = \left[ \sum d_{kl}^k(s) c_k t_k^k \right]^{-1} \sum d_{kl}^k(s) c_k t_k^k t_{lT}^k(\tilde{\theta}),$$

$$J(\tilde{\theta}) = - \sum d_{kl}^k(s) g_k(d(s)) (\partial l^k(\tilde{\theta}) / \partial \theta^T),$$

et  $I^p$  est la matrice identité  $p \times p$ .

Après certaines simplifications, la première composante  $\text{est}(I)$  est donnée par (3.2) ou (3.3) avec  $b_k$  transformé en

$$Z_k h^k = [J(\tilde{\theta})]^{-1} e_k(\tilde{\theta}) g_k(d(s)), \quad (3.5)$$

où

$$e_k(\tilde{\theta}) = l^k(\tilde{\theta}) - b_T^l t_k.$$

De même, la deuxième composante  $\text{est}(II)$  se réduit à

$$\text{est}(II) = [J(\tilde{\theta})]^{-1} \sum d_{kl}^k(s) g_k^2(d(s)) l^k(\tilde{\theta}) l^T(\tilde{\theta}) [J(\tilde{\theta})]^{-1}. \quad (3.6)$$

si  $\text{Cov}_m[l^k(\tilde{\theta}) l^T(\tilde{\theta})] = 0$  pour  $k \neq l$ . L'estimateur de la variance totale de  $\tilde{\theta}$  est maintenant estimé par

Nous avons effectué une étude en simulation pour comparer les propriétés relatives des trois estimateurs de variance  $g^{\text{DR}}$ ,  $g^{\text{cus}}$  et  $g^{\text{mix}}$ , dans le cas particulier d'un modèle de régression logistique :

$$E_m(y^k) = \mu^k(\tilde{\theta}) = \exp(x_T^k \tilde{\theta}) / [1 + \exp(x_T^k \tilde{\theta})] \quad (3.8)$$

$$V_m(y^k) = \mu^k(\tilde{\theta}) (1 - \mu^k(\tilde{\theta})), \text{Cov}_m(y^k, y^l) = 0, k \neq l.$$

Dans ces conditions, nous avons  $l^k(\tilde{\theta}) = x^k(y^k - \mu^k(\tilde{\theta}))$ , et

$$J(\tilde{\theta}) = \sum d_{kl}^k(s) g_k(d(s)) x^k x_T^k \mu^k(\tilde{\theta}) (1 - \mu^k(\tilde{\theta})).$$

Pour l'étude en simulation, nous posons que  $x^k =$

$(1, x_k)^T$ , où les  $x_k$  désignent le nombre de lits pour la population d'hôpitaux de taille  $N = 393$  étudiée à la section 2.2. Nous avons exécuté une poststratification en divisant la population en deux classes, avec  $N_1 = 171$  hôpitaux  $k$  ayant  $x_k < 350$  dans la classe 1 et  $N_2 = 122$  hôpitaux  $k$  avec  $x_k \geq 350$  dans la classe 2. Ici,  $g_k(d(s)) = N_h / N^h$ , si  $k$  appartient à la classe  $h$ , où  $N^h = \sum d_{kl}^k(s) t_{hk}$  est l'estimateur pondéré par les poids de sondage de  $N^h$ , et  $t_k = (t_{1k}, t_{2k})^T$  est le vecteur de variables indicatrices de classe  $t_{hk}$ .

Nous avons généré  $R = 40\,000$  populations finies  $\{y_1^k, \dots, y_N^k\}$ , chacune de taille  $N = 393$ , en émettant l'hypothèse du modèle de régression logistique (3.8) avec  $\theta = (\theta_0, \theta_1)^T = (-1, 0.005)^T$ . Le paramètre d'intérêt est  $\theta_1 = 0.005$ . Dans chacune des populations produites, nous avons sélectionné un échantillon aléatoire simple de taille  $n = 150$ , puis obtenu l'estimateur  $\hat{\theta}_1$  estimé et pondéré par calage et les estimateurs de variance connexes  $\text{est}(I) = g_s(\hat{\theta}_1)$ ,  $g^{\text{DR}}(\hat{\theta}_1)$  et  $g^{\text{mix}}(\hat{\theta}_1)$  à partir de chaque échantillon  $r$ . Nous avons obtenu les moyennes des estimations et des estimations de variance comme étant  $\text{moy}(\hat{\theta}_1) \approx 0.00514$ ,  $\text{moy}(g^{\text{DR}}) \approx 0.0989$ ,  $\text{moy}(g^{\text{cus}}) \approx 0.0987$ ,  $\text{moy}(g^{\text{mix}}) \approx 0.0988$  et  $\text{moy}(g_s) \approx 0.0613$ . En outre, l'EQM totale estimée de  $\hat{\theta}_1$  est égale à 0.0998. Donc, conditionnellement, l'estimateur  $\hat{\theta}_1$  est approximativement sans biais pour  $\theta_1$ , et le biais des trois estimateurs de variance  $g^{\text{DR}}$ ,  $g^{\text{cus}}$  et  $g^{\text{mix}}$  est négligeable. Par ailleurs, ignorer la deuxième composante et utiliser seulement la



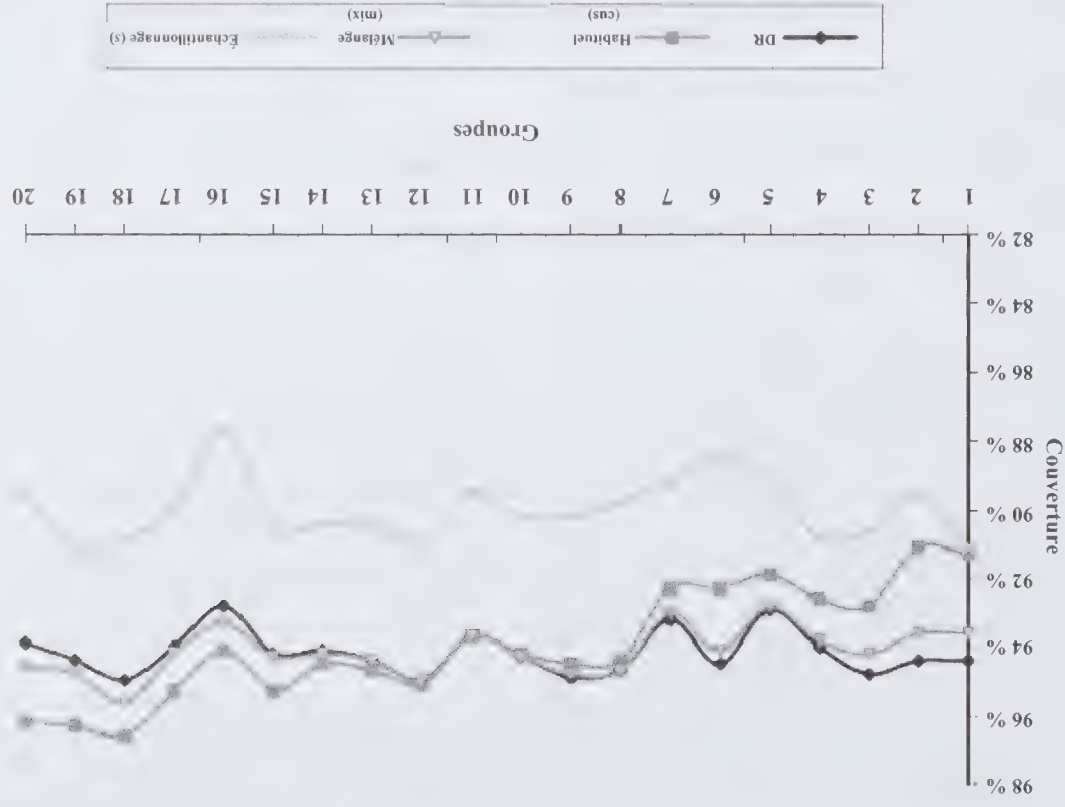


Figure 4 Taux de couverture conditionnels des intervalles de confiance de 95 % : modèle de ratio  
 $9_{DR}$ ,  $9_{cus}$ ,  $9_{mix}$  et  $9_s$  pour le niveau nominal de confiance de 95 % : modèle de ratio

### 3. Équations d'estimation pondérées par calage

#### 3.1 Estimateurs des paramètres du modèle

Supposons que le modèle de superpopulation appliqué aux réponses  $y_k$  est spécifié par un modèle linéaire généralisé (McCullagh et Nelder 1989) de moyenne  $E_m(y_k) = \mu_k(\theta) = h(x_k^T \theta)$ , où  $x_k$  est un vecteur  $p \times 1$  de variables explicatives,  $\theta$  est le vecteur  $p$  des paramètres du modèle et  $h(\cdot)$  est une fonction « lien ». Par exemple,  $h(a) = a$  donne un modèle de régression linéaire et  $h(a) = e^a / (1 + e^a)$ , donne un modèle de régression logistique pour les réponses binaires  $y_k$ .

Nous définissons des équations d'estimation en population finie (EBPF), basées sur les fonctions d'estimation  $l_k(\theta)$ , comme étant  $l(\theta) = \sum l_k(\theta) = 0$  avec  $E_m l_k(\theta) = 0$ , et la solution de ces EBPF donne le vecteur de paramètres de population finie  $\theta_N$ . Par exemple,  $l_k(\theta) = x_k(y_k - \mu_k(\theta))$  pour les modèles de régression linéaire et logistique. Nous utilisons les poids de régression généralisée (GREG)  $w_k(s) = d_k(s)g_k(d(s))$ , où les poids  $g$  sont donnés par  $g_k(d(s)) = 1 + (T - T)^T [\sum d_k(s) c_k t_k t_k^T]^{-1} c_k t_k$

#### 3.2 Estimateurs de variance linéarisés

Nous commençons par étendre le résultat de l'estimation de la variance pour le cas scalaire  $U = \sum b_k^T d_k$  (section 2.2)

La solution de (3.1), obtenue par la méthode itérative de type Newton-Raphson, donne l'estimateur pondéré par calage  $\hat{\theta}$  de  $\theta$ , et  $\hat{\theta}$  est approximativement sans biais sous le plan et le modèle pour  $\theta$ , c'est-à-dire  $E(\hat{\theta}) \approx \theta$ . Il découle de (3.1) que  $\hat{\theta}$  est de la forme  $f(A_d)$  avec  $d_k = (d_k(s), d_k(s)l_k^T(\theta))^T$ , où  $f(A_d)$  est un vecteur  $p \times 1$  et  $A_d$  est une matrice  $(p+1) \times N$  dont la  $k^e$  colonne est  $d_k$ . Ici, nous avons  $h_k = 1$  et  $(h_{2k}, \dots, h_{(p+1)k}) = l_k(\theta)$ .

sont donnés par

$$l(\theta) = \sum w_k(s) l_k(\theta) = \sum d_k(s) g_k(d(s)) l_k(\theta) = 0. \quad (3.1)$$

Nous utilisons les poids de calage,  $w_k(s)$ , pour estimer les EBPF. Les équations d'estimation pondérée par calage des totaux  $Y = \sum y_k$ , quand la relation entre  $y_k$  et  $t_k$  est linéaire (Sæmndal, Swensson et Wretman 1989, chapitre 6).  $T$  et produisent des estimateurs efficaces  $\hat{Y} = \sum w_k(s) y_k$  et  $\hat{T} = \sum w_k(s) t_k$ , ont la propriété de calage  $\sum w_k(s) t_k = \sum d_k(s) t_k$ . Les poids GREG,  $w_k(s)$ , ont la propriété de calage  $\sum w_k(s) t_k = \sum d_k(s) t_k$  et  $d(s)$  est le vecteur  $N \times 1$  des poids  $d_k(s)$ . Les HT du total connu  $T$  d'un vecteur  $q \times 1$  de variables de calage  $t_k$  et  $d(s)$  sont spécifiés, où  $T = \sum d_k(s) t_k$  est l'estimateur

L'utilisation de  $\theta_s$  donne lieu à un défaut de couverture important, parce que la fraction d'échantillonnage, 100/393, est grande. Par ailleurs, le TCC associé à  $\theta_{DR}$  est plus proche du niveau nominal dans les divers groupes, tandis que  $\theta_{cus}$  présente une tendance à travers les groupes, le TCC variant de 91 % à 97 %. De surcroît, le TCC associé à  $\theta_{mix}$  est légèrement inférieur à celui associé à  $\theta_{DR}$  pour la première moitié des groupes, mais  $\theta_{mix}$  et  $\theta_{DR}$  donnent des résultats comparables.



Figure 2 Biases relatives conditionnelles de l'estimateur par facteur d'extension et de l'estimateur par le ratio : modèle de ratio

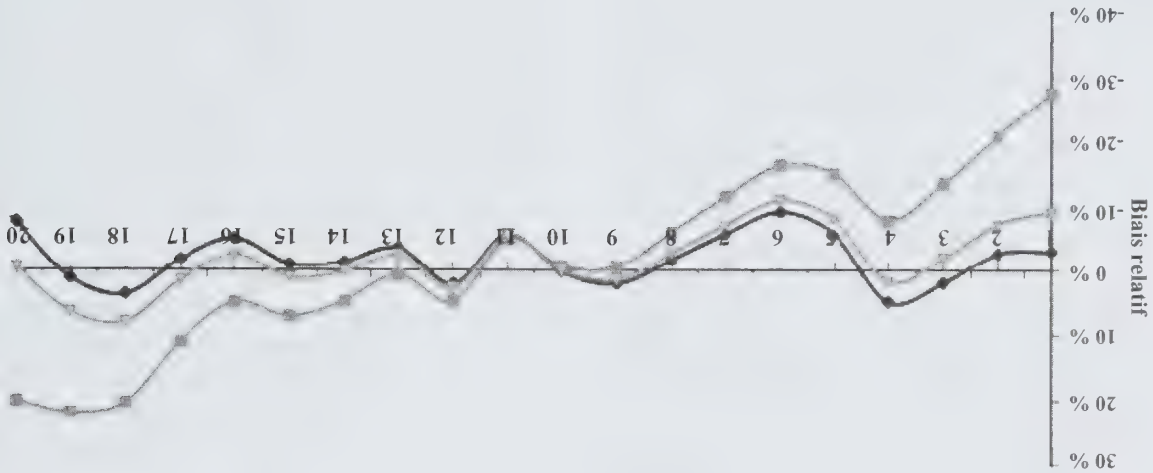


Figure 3 Biases relatives conditionnelles des estimateurs de variance de variance  $\theta_{DR}$ ,  $\theta_{cus}$  et  $\theta_{mix}$  : modèle de ratio

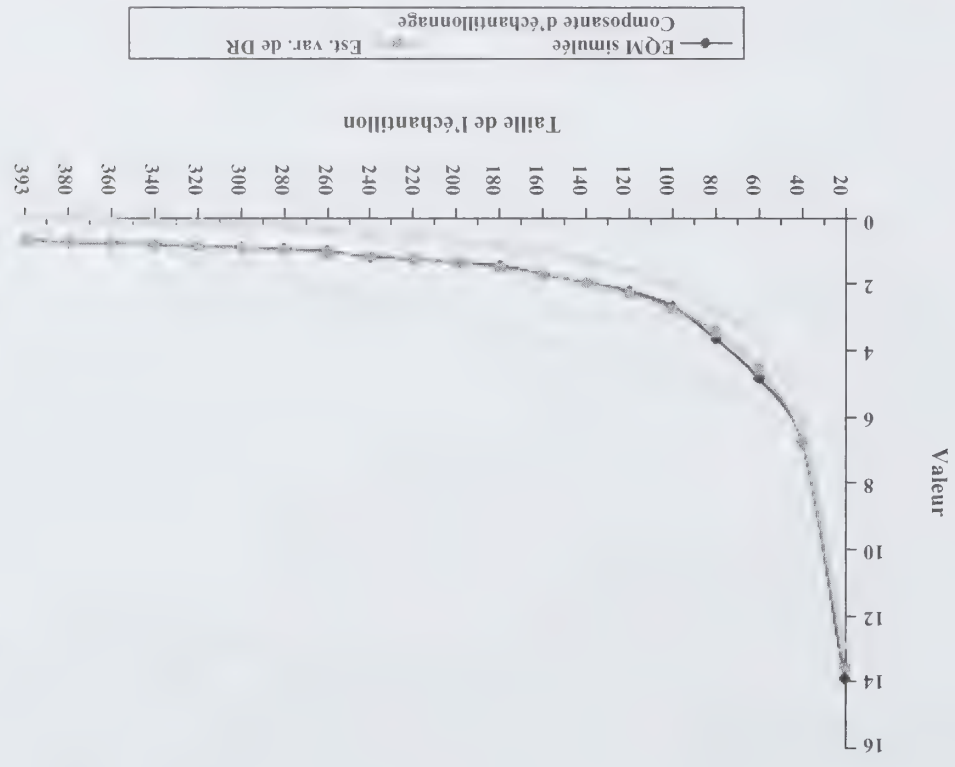


Figure 1 Moyennes des estimations de variance pour certaines tailles d'échantillons comparativement à l'EOM estimée de l'estimateur par le ratio.  $\theta_{DR}$  = Est. var. de DR,  $\theta_s$  = Composante d'échantillonnage : modèle de ratio

Nous avons également examiné les propriétés conditionnelles des estimateurs de variance sous échantillonnage aléatoire simple sachant  $\bar{x}$ , au moyen d'une autre étude en simulation pour l'inférence au sujet de  $\theta$ , en utilisant le modèle (2.15). L'étude est semblable à celle de Royall et Cumberland (1981) pour l'inférence au sujet de la moyenne de population finie  $\theta_N = \bar{Y}$  à partir d'une population fixe  $\{Y_1, \dots, Y_N\}$ . Nous avons généré  $R = 20\,000$  populations finies  $\{Y_1, \dots, Y_N\}$ , chacune de taille  $N = 393$ , à partir de (2.15) en utilisant le nombre de lits comme  $x_k$  et, pour chaque population, nous avons ensuite sélectionné un échantillon aléatoire simple de taille  $n = 100$ . Nous avons classé les 20 000 échantillons par ordre croissant de valeur de  $\bar{x}$ , puis nous les avons regroupés en 20 groupes, chacun de taille 1 000, de façon que le premier groupe,  $G_1$ , contienne les 1 000 échantillons ayant les plus petites valeurs de  $\bar{x}$ , que le groupe suivant,  $G_2$ , contienne les 1 000 plus petites valeurs de  $\bar{x}$  suivantes, et ainsi de suite pour obtenir  $G_1, \dots, G_{20}$ . Pour chacun des 20 groupes ainsi formés, nous avons calculé la moyenne des estimations du ratio  $\hat{\theta} = \bar{X}(\bar{y}/\bar{x})$  et l'estimation moyenne  $\bar{y}$ , ainsi que le biais relatif conditionnel (BRC) résultant en estimant  $\theta = 2\bar{X}$ ; voir la figure 2. Il est clair, si l'on examine cette figure, que  $\bar{y}$  est conditionnellement biaisé, contrairement à

$\theta$  : BRC négatif (-14 %) pour  $G_1$  augmentant pour passer à un BRC positif (+14 %) pour  $G_{20}$ . Notons que  $\bar{y}$  et  $\hat{\theta}$  sont tous deux inconditionnellement sans biais pour  $\theta$ . Le biais conditionnel de  $\hat{\theta}$  et de  $\bar{y}$  lorsque nous estimons le paramètre du modèle  $\theta$  est semblable au biais conditionnel dans l'estimation du paramètre de population finie  $\theta_N = \bar{Y}$ , comme l'ont observé Royall et Cumberland (1981). Nous avons également calculé l'EQM conditionnelle de  $\hat{\theta}$  et le BRC associé des estimateurs de variance  $\theta_{DR}$ ,  $\theta_{cus}$  et  $\theta_{mix}$  basés sur les valeurs moyennes de  $\theta_{DR}$ ,  $\theta_{cus}$  et  $\theta_{mix}$  dans chaque groupe; voir la figure 3. Il est évident, si l'on examine cette figure, que le BRC de  $\theta_{cus}$  varie de -28 % à 20 % lorsque l'on passe d'un groupe à l'autre, alors que  $\theta_{DR}$  ne manifeste pas ce genre de tendance et que son BRC est inférieur à 5 % en valeur absolue, sauf pour  $G_6$  et  $G_{20}$ . En outre, le BRC de  $\theta_{mix}$  est en grande partie négatif et inférieur à celui de  $\theta_{DR}$  pour la première moitié des groupes et supérieur pour la deuxième moitié, mais  $\theta_{mix}$  ne présente aucune tendance discernable, contrairement à  $\theta_{cus}$ . La figure 4 donne les taux de couverture conditionnelle (TCC) des intervalles de confiance de la théorie normale fondés sur  $\theta_{DR}$ ,  $\theta_{cus}$ ,  $\theta_{mix}$  et  $\theta_s$  (en ignorant la composante  $\theta_m$ ) au niveau nominal de 95 %. Comme prévu,



en utilisant (2.9). De plus, en remplaçant  $\mathbf{u}_k$  par  $\mathbf{z}_k$  dans (2.6), nous obtenons

$$\mathbf{z}_T^T \text{cov}_m(\mathbf{h}_k, \mathbf{h}_l) \mathbf{z}_l = \mathbf{z}_{2k}^T \mathbf{z}_{2l}^T \text{cov}_m(\mathbf{y}_k, \mathbf{y}_l).$$

Sous le modèle de ratio (2.1) avec la variance sous le modèle  $V^m(\mathbf{y}_k) = \sigma_k^2 = E^m(\mathbf{y}_k - \beta \mathbf{x}_k)^2$  par  $(\mathbf{y}_k - R \mathbf{x}_k)^2$  et pouvons estimer  $\sigma_k^2 = E^m(\mathbf{y}_k - \beta \mathbf{x}_k)^2$  par  $(\mathbf{y}_k - R \mathbf{x}_k)^2$  et poser que  $\text{cov}_m(\mathbf{y}_k, \mathbf{y}_l) = 0$ , pour  $k \neq l$ .

Nous étudions maintenant le cas particulier de l'échantillonnage aléatoire simple sans remise. Dans ce cas, aussi bien (2.4) que (2.5) se réduit à

$$\text{est}(I) = \left( \frac{\bar{X}}{\bar{X}} \right)^2 \left( 1 - \frac{n}{N} \right) s_e^2, \quad (2.10)$$

où  $s_e^2 = \sum a_k(s) e_k^2 / (n-1)$ , et (2.6) se réduit à

$$\text{est}(II) = \left( \frac{\bar{X}}{\bar{X}} \right)^2 \frac{n}{(n-1)} s_e^2. \quad (2.11)$$

D'où, en utilisant (2.10) et (2.11), l'estimateur de variance (2.8) se réduit à

$$\mathfrak{g}^{\text{DR}}(\theta) = \text{est}(I) + \text{est}(II)$$

$$= \left( \frac{\bar{X}}{\bar{X}} \right)^2 \frac{n}{1} \frac{1}{N-1} s_e^2. \quad (2.12)$$

Il est intéressant de noter que le « poids  $\bar{X} / \bar{x}$  apparaît automatiquement dans  $\mathfrak{g}^{\text{DR}}(\theta)$ , donné par (2.12), et que la correction pour population finie  $1 - n/N$  est absente dans  $\mathfrak{g}^{\text{DR}}(\theta)$  contrairement à  $\text{est}(I)$  donné par (2.10).

Suivant l'approche habituelle de l'estimation de la variance totale (voir, par exemple, Korn et Graubard 1998),  $V(\theta)$  s'écrit d'abord

$$V(\theta) = E^m V^p(\theta) + V^m E^p(\theta)$$

$$\approx E^m V^p(\theta) + V^m(\bar{Y})$$

$$= E^m V^p(\theta) + N^{-2} \sum E^m(\mathbf{y}_k - \beta \mathbf{x}_k)^2, \quad (2.13)$$

sous le modèle du ratio avec  $\sigma_k^2$ ,  $k = 1, \dots, N$  non spécifié. Le premier terme  $E^m V^p(\theta)$  de (2.13) est alors estimé au moyen d'un estimateur convergent sous le plan de  $V^p(\theta)$ , habituellement (2.10) sans le facteur  $g(\bar{X}/\bar{x})^2$ . Le deuxième terme est estimé par  $N^{-2} \sum d_k(s) (\mathbf{y}_k - R \mathbf{x}_k)^2 = (nN)^{-1} (n-1) s_e^2$ . La somme des deux termes est alors égale à (2.12) sans le facteur  $g$ . Nous désignons cet estimateur de variance habituel par  $\mathfrak{g}^{\text{cus}}(\theta)$ . Par ailleurs, si (2.10) avec le facteur  $g$  est utilisé pour estimer  $V^p(\theta)$ ,

la somme de ce terme estimé et de l'estimateur précédent du deuxième terme mène à un estimateur de variance « hybride »

$$\mathfrak{g}^{\text{mix}}(\theta) = \text{est}(I) + (nN)^{-1} (n-1) s_e^2$$

où le terme  $g$  est absent dans le dernier terme. Les résultats qui précèdent montrent clairement que le choix de l'estimateur de la variance totale sous l'approche habituelle n'est pas unique, contrairement à la situation sous l'approche proposée. Si le paramètre d'intérêt est  $\beta = \theta / \bar{X}$  au lieu de  $\theta$ , alors  $\beta = \theta / \bar{X} = R$  et  $\mathfrak{g}^{\text{DR}}(\beta)$  sous l'échantillonnage aléatoire simple est donné par

$$\mathfrak{g}^{\text{DR}}(\beta) = \bar{X}^{-2} \mathfrak{g}^{\text{DR}}(\theta) = \bar{X}^{-2} \frac{n}{1} \frac{1}{N-1} s_e^2. \quad (2.14)$$

L'approche habituelle aboutit au même estimateur de variance (2.14).

## 2.4 Étude en simulation

Nous avons effectué une petite étude en simulation pour examiner les propriétés des divers estimateurs de variance, conditionnellement et inconditionnellement à  $\bar{X}$ . Nous avons d'abord généré  $R = 2000$  populations finies  $\{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ , chacune de taille  $N = 393$ , au moyen du modèle de ratio

$$\mathbf{y}_k = 2 \mathbf{x}_k + \mathbf{x}_k^{1/2} \mathbf{e}_k, \quad (2.15)$$

avec les valeurs indépendantes  $\mathbf{e}_k$ , tirées de  $N(0, 1)$ , où les  $\mathbf{x}_k$  fixes sont les « nombres de lits » pour la population d'hôpitaux étudiée dans Valliant, Dorfman et Royall (2000, page 424-427). Un échantillon aléatoire simple de taille spécifiée  $n$  est tiré de chaque population générée. Notre paramètre d'intérêt est  $\theta = \beta \bar{X}$ , où  $\beta = 2$ .

L'EQM totale simulée de l'estimateur par le ratio  $\theta = \bar{X}(\bar{y}/\bar{x})$  est calculée comme  $M(\theta) = R^{-1} \sum_{r=1}^{2000} (\hat{\theta}_r - \theta)^2$ , où  $\hat{\theta}_r$  est la valeur de  $\theta$  pour le  $r^{\text{e}}$  échantillon simulé et  $(\bar{y}, \bar{x})$  sont les moyennes d'échantillon. Nous avons estimé la variance totale  $\mathfrak{g}^{\text{DR}}(\theta)$ , et ses composantes  $\mathfrak{g}^s = \text{est}(I)$  et  $\mathfrak{g}^m = \text{est}(II)$  à partir de chaque échantillon simulé  $r$ , ainsi que leurs moyennes  $\bar{\mathfrak{g}}^{\text{DR}}$ ,  $\bar{\mathfrak{g}}^s$  et  $\bar{\mathfrak{g}}^m$  sur  $r$ . La figure 1 représente graphiquement la moyenne des estimations de la variance,  $\bar{\mathfrak{g}}^{\text{DR}}$  et  $\bar{\mathfrak{g}}^s$ , ainsi que l'EQM totale simulée pour  $n = 20, 40, \dots, 380, 393$ . Dans le cas où  $n = N$ ,  $\bar{\mathfrak{g}}^s = 0$ . L'examen de la figure 1 montre que  $\bar{\mathfrak{g}}^{\text{DR}}$  est approximativement sans biais, tandis que  $\bar{\mathfrak{g}}^s$  donne lieu à une sous-estimation grave à mesure que la taille d'échantillon,  $n$ , augmente.

respectivement, et  $\sigma^2 > 0$ . Supposons que nous voulons estimer la moyenne de superpopulation  $\theta = E_m(\bar{Y}) = N^{-1} \sum E_m(y_k) = \beta \bar{X}$ , où  $\bar{X}$  est la moyenne de population finie de  $y$ . Dans ce cas, un estimateur par le ratio de  $\theta$  est donné par

$$\hat{\theta} = \bar{X}(\bar{Y}/\bar{X}) \equiv \bar{X}\hat{R}, \quad (2.2)$$

où  $\bar{Y} = \sum d_k(s)y_k$  et  $\bar{X} = \sum d_k(s)x_k$  sont les estimateurs sans biais sous le plan des totaux  $Y$  et  $X$ , et  $\bar{X}$  est la moyenne de population connue de  $x$ . Nous pouvons écrire l'estimateur par le ratio (2.2) sous la forme  $\hat{\theta} = \bar{X}(\sum d_{2k} / \sum d_{1k} x_k)$ , où  $d_{1k} = d_k(s)$  et  $d_{2k} = d_k(s)y_k$ . Il s'agit d'un cas particulier de  $f(A_d)$  avec  $p = 1$  et  $h_{2k} = y_k$ .

Soit  $E_p$  l'espérance sous le plan et  $E = E_m E_p$ , l'espérance totale. Alors, nous avons  $E(d_{1k}^i) = E_m(1) = 1 = \mu_{1k}$  et  $E(d_{1k}^{ik}) = E_m(g_{ik}^i) \equiv \mu_{ik}$ ,  $i = 2, \dots, p + 1$ , en notant que  $E_p(d_k^i(s)) = 1$ . Nous supposons que  $f(A_u) = \theta$ , où  $A_u$  est une matrice  $(p + 1) \times N$  avec les colonnes  $\mu_k = (\mu_{1k}, \mu_{2k}, \dots, \mu_{(p+1)k})^T$ . Donc,  $\theta$  est asymptotiquement sans biais sous le plan  $p$  et sous le modèle  $m$  pour  $\theta$ . Dans le cas particulier de l'estimateur par le ratio, nous avons  $f(A_u) = \beta \bar{X} = \theta$ , en notant que  $\mu_{1k} = 1$  et  $\mu_{2k} = \beta x_k$ .

## 2.2 Estimateur de variance par linéarisation

Nous commençons par dériver un estimateur de la variance totale d'un estimateur linéaire  $\bar{U} = \sum u_k^T d_k$ , où  $u_k$  est un vecteur de constantes. La variance totale de  $\bar{U}$  peut être décomposée comme il suit

$$V(U) = E_m V_p(U) + V_m E_p(U) \equiv I + II, \quad (2.3)$$

où  $V_p$  et  $V_m$  désignent la variance sous le plan et la variance sous le modèle, respectivement. Un estimateur sans biais sous le plan de la composante  $I$  de la variance totale (2.3) s'obtient en estimant la variance sous le plan  $V_p(U)$  pour des  $h_k = (h_{1k}, \dots, h_{(p+1)k})^T$  fixes. Maintenant, en notant que  $\bar{U} = \sum b_k d_k(s)$  est l'estimateur classique de Narain-Horvitz-Thompson (NHT) du total  $U = \sum b_k$  quand les  $b_k = u_k^T h_k$  sont fixes conditionnellement, nous pouvons utiliser soit l'estimateur de variance de Sen-Yates-Grandy (SYG) pour des plans avec tailles d'échantillon fixes ou l'estimateur de variance d'Horvitz-Thompson (HT) pour des plans arbitraires. L'estimateur SYG est donné par

$$\text{est}(I) = g_{\text{SYG}}(U)$$

$$= \sum \sum_{k < l} d_{kl}(s) \frac{\pi_k \pi_l}{(\pi_k \pi_l - \pi_{kl})} (b_k - b_l)^2, \quad (2.4)$$

$$z_k^T h_k = z_{1k} + z_{2k} y_k = (\bar{X}/\bar{X})(y_k - \hat{R} x_k) \equiv (\bar{X}/\bar{X}) e_k,$$

En outre, dans (2.4) ou (2.5),  $b_k$  est remplacé par

$$z_k = (\bar{X}/\bar{X})(-R x_k, 1)^T = (z_{1k}, z_{2k})^T. \quad (2.9)$$

Pour l'estimateur par le ratio  $\hat{\theta} = \bar{X}\hat{R}$  du paramètre du

## 2.3 Cas particulier de l'estimateur par le ratio

que nous obtenons à partir de  $g(u)$  en remplaçant  $u_k$  par la « variable linéarisée »  $z_k = \partial f(A_b) / \partial b_k |_{A_b = A_u}$ . Une justification théorique rigoureuse de (2.8) ressemble à celle de Deville (1999).

$$g_{\text{DR}}(\theta) = g(z), \quad (2.8)$$

sous la forme

où  $z_k = \partial f(A_b) / \partial b_k |_{A_b = A_u}$  et  $A_b$  est une matrice  $(p + 1) \times N$  dont la  $k^{\text{e}}$  colonne  $b_k$  est un vecteur de nombres réels arbitraires. L'approximation (2.7) est valide pour tout  $\theta$  qui peut être exprimé sous forme d'une fonction lisse des totaux estimés. En suivant l'exemple de Denuit et Rao (2004), nous pouvons maintenant écrire un estimateur par linéarisation de la variance totale sous la forme

$$\hat{\theta} - \theta \approx \sum z_k^T (d_k - \mu_k) \quad (2.7)$$

$\hat{\theta} - \theta$  sous la forme

Penchons-nous maintenant sur l'estimation de la variance totale de  $\hat{\theta}$ . À l'instar de Denuit et Rao (2004), nous pouvons écrire un développement en série de Taylor de

après avoir remplacé  $\text{Cov}_m(h_k, h_l)$  par un estimateur  $\text{cov}_m(h_k, h_l)$ . L'estimateur de la variance totale (2.3) est maintenant donné par  $\text{est}(I) + \text{est}(II)$ . Nous le désignons,

$$\text{est}(II) = \sum \sum d_{kl}(s) u_k^T \text{cov}_m(h_k, h_l) u_l, \quad (2.6)$$

et  $m$  est alors donné par

Si nous nous tournons vers la composante  $II$  de la variance totale (2.3), nous avons  $V_m E_p(U) = V_m(\sum u_k^T h_k) = \sum \sum u_k^T \text{Cov}_m(h_k, h_l) u_l$ , et un estimateur sans biais sous  $p$

où  $d_{kk}(s) = d_k(s)$ . Pour le cas particulier de l'échantillon-

$$\text{est}(I) = g_{\text{HT}}(U) = \sum \sum d_{kl}(s) \frac{\pi_k \pi_l}{(\pi_k \pi_l - \pi_{kl})} b_k b_l, \quad (2.5)$$

HT est donné par

où  $d_{kl}(s) = \{a_k(s)a_l(s)\} / \pi_{kl}$  et  $\pi_{kl}$  est la probabilité d'inclusion des unités  $k$  et  $l$  ( $k \neq l$ ). L'estimateur de variance



$$A(\theta) = E(A^d(\theta) + A^w(\theta) + A^N(\theta)), \quad (1.1)$$

$$(1.1) \quad (\theta^N)^w A + (\theta)^d A^w E = (\theta) A$$

(1966). Dans l'analyse des données d'enquête, on suppose fréquemment que les valeurs de population  $y_k, k = 1, \dots, N$ , sont issues d'un modèle de superpopulation, et l'utilisateuse cherche souvent à faire des inférences au sujet des paramètres du modèle. Soit  $\theta^N$  un paramètre de population finie, c'est-à-dire un estimateur d'un paramètre du modèle  $\theta$  quand les valeurs de population  $y_k$  sont toutes connues, et soit  $\hat{\theta}$  un estimateur sans biais sous le plan de  $\theta^N$ , le paramètre de population finie. Supposons que  $\hat{\theta}$  est sans biais sous le plan et sous le modèle pour  $\theta$ , c'est-à-dire que  $E^m E^p(\hat{\theta}) = \theta$ , où  $E^m$  et  $E^p$  désignent les espérances sous le plan et sous le modèle, respectivement. Alors, la variance totale de  $\hat{\theta}$  est  $V(\hat{\theta}) = E^m E^p(\hat{\theta}) - (\theta)^2$ , qui peut être décomposée comme il suit

arbitraire.

Dernati et Rao (2004) ont étudié des estimateurs généraux qui peuvent être exprimés sous forme de fonctions lissées des poids  $\mathbf{d}(s) = \{d_1(s), \dots, d_N(s)\}^T$ , disons  $\hat{\theta} = f(\mathbf{d}(s))$ , et ont obtenu un estimateur de variance par linéarisation de Taylor directement sous la forme  $\vartheta(z)$  avec des variables linéarisées connues  $z_k = \partial f(b) / \partial b^k|_{b=d(s)}$  sans estimer la variable  $z_k$  au préalable, puis la remplacer par un estimateur. Par exemple, dans le cas de l'estimateur par le ratio, leur méthode mène automatiquement à  $z_k$  susmentionné. Cette méthode peut être appliquée à divers estimateurs, y compris ceux des paramètres de régression logistique en population finie fondés sur les poids de calage (Dernati et Rao 2004). Les travaux antérieurs sur l'estimation directe de la variance comprennent ceux de Binder

estimateur  $z_k$  qui peut être fondé sur la méthode de substitution. Deville (1999) a dérivé un estimateur de variance par linéarisation de Taylor de la fonctionnelle  $T(M)$  de la forme  $\mathfrak{g}(z)$ , où  $z_k = I^T(M; y^k)$  désigne la fonction d'influence de  $T$  à la valeur  $y^k$ , puis a remplacé  $z_k$  par l'estimateur d'échantillon  $z_{k1} = I^T(\hat{M}; y^k)$ . Par exemple, quand  $\theta$  est l'estimateur par le ratio  $(Y/X)X = \bar{R}X$  du total  $Y$ , où  $X = Y(x)$  et  $X = Y(x)$  est le total connu d'une variable auxiliaire  $x$ , nous obtenons  $\hat{z}_k = y^k - R x^k$  et  $z_{k1} = y^k - R x^k$ . Cependant,  $z_k = (X/X)(y^k - R x^k)$  est aussi un candidat pour estimer  $z_k$  et l'estimateur de variance résultant  $\mathfrak{g}(z)$  est souvent préféré à  $\mathfrak{g}(z_1)$ ; voir Demnati et Rao (2004). Donc, sous l'approche de Deville, le choix d'un estimateur de  $z_k$  est dans une certaine mesure

covariables  $x_k$  fixes :

$$E_m(y_k) = \beta x_k, V_m(y_k) = \sigma^2 x_k, \text{COV}_m(y_k, y_l) = 0, \quad k \neq l, k, l = 1, \dots, N, \quad (2.1)$$

$$(2.1) \quad k \neq t, k, t = 1, \dots, N,$$

Considérons une population finie  $U$  de  $N$  éléments, et soit  $d^k(s) = a^k(s)/\pi_k$  le poids de sondage attaché à l'élément de population  $k$ , où  $a^k(s) = 1$  si l'élément  $k$  est compris dans l'échantillon  $s$  et  $a^k(s) = 0$  autrement, et  $\pi_k$  est la probabilité d'inclusion associée à l'élément  $k$ . Nous considérons les estimateurs  $\theta$  d'un paramètre scalaire  $\theta$  qui peuvent être exprimés comme des fonctions de variables aléatoires sous le plan et le modèle supposé. En particulier,  $\theta = f(\mathbf{A})$ , où  $\mathbf{A}$  est une matrice  $(p+1) \times N$  avec les colonnes  $\mathbf{d}^k = (d^k h^k, d^k h^{2k}, \dots, d^k h^{(p+1)k})^T \equiv (d^{1k}, \dots, d^{(p+1)k})^T$  où  $d^k = d^k(s)$  est aléatoire sous le plan,  $h^{ik} = 1$  et  $h^{ik} = 0$  si  $i = 1, \dots, p$  et  $k = 1, \dots, N$ . Par exemple, considérons le modèle du ratio avec covariables  $x_k$  fixes :

## 2.1 Estimateurs ponctuels

## 2. Paramètre du modèle scalaire

À la section 2, nous considérons le cas d'un paramètre scalaire  $\theta$  et présentons des estimateurs de variance par linéarisation, en élargissant l'approche de Demnat et Rao (2004). Nous illustrons la méthode pour le cas particulier d'un estimateur par le ratio d'une moyenne de superpopulation  $\theta$ . À la section 3, nous étendons les résultats de la section 2 aux estimateurs d'un paramètre vectoriel  $\theta$  dont les valeurs sont les solutions d'équations d'estimation pondérées et nous illustrons la méthode pour le cas particulier des paramètres d'un modèle de régression logistique. Nous présentons aussi les résultats de simulations.

(Skinner, Holt et Smith 1989, page 14). Par ailleurs, il est nécessaire d'estimer la variance totale  $V(\theta)$  quand la variance sous le modèle  $W(\theta^N)$  n'est pas négligeable comparativement à  $E^m V^d(\theta)$ . Il faut pour cela prendre en considération conjointement les processus aléatoires sous le plan et sous le modèle. Molina, Smith et Sugden (2001) soutiennent que le processus combiné de génération de la population finie et de sélection de l'échantillon devrait servir de fondement aux inférences analytiques concernant les paramètres du modèle. Rubin-Bleuer et Schiopu-Kratina (2005) ont donné un cadre mathématique pour l'inférence conjointe sous le modèle et sous le plan. Cependant, une méthode dont l'application est générale est nécessaire pour l'estimation de la variance totale. Le principal objectif du présent article est de proposer une telle méthode, en étendant l'approche de Demnati-Rao aux paramètres de



# Estimateurs de variance par linéarisation pour les paramètres de modèles à partir de données d'enquêtes complexes

Abdellatif Demnati et J.N.K. Rao<sup>1</sup>

## Résumé

Les méthodes de linéarisation de Taylor sont souvent utilisées pour obtenir des estimateurs de la variance d'estimateurs par calage de totaux et de paramètres de population finie (ou de recensement) non linéaires, tels que des ratios ou des coefficients de régression et de corrélation, qui peuvent être exprimés sous forme de fonctions lisses de totaux. La linéarisation de Taylor s'applique généralement à tout plan d'échantillonnage, mais elle peut produire de multiples estimateurs de la variance qui sont asymptotiquement sans biais par rapport au plan en cas d'échantillonnage répété. Le choix parmi les estimateurs de variance doit donc s'appuyer sur d'autres critères, tels que i) l'absence approximative de biais dans la variance par rapport au modèle de l'estimateur obtenu sous un modèle hypothétique et ii) la validité sous variance par linéarisation de Taylor produisant directement un estimateur de variance unique qui satisfait aux critères susmentionnés pour des plans de sondage généraux. Dans l'analyse des données d'enquête, on suppose généralement que les populations finies sont générées au moyen de modèles de superpopulation et l'on s'intéresse aux inférences analytiques concernant les paramètres de ces modèles. Si les fractions d'échantillonnage sont faibles, la variance d'échantillonnage reflète presque toute la variation due aux processus aléatoires liés au plan de sondage et au modèle. Par contre, si les fractions d'échantillonnage ne sont pas négligeables, il faut tenir compte de la variance du modèle pour construire des inférences valides concernant les paramètres du modèle sous le processus combiné de génération de la population finie à partir du modèle hypothétique de superpopulation et de sélection de l'échantillon conformément au plan de l'échantillonnage spécifique. Dans le présent article, nous obtenons un estimateur de la variance totale selon l'approche de Demnati-Rao en supposant que les caractéristiques d'intérêt sont des variables aléatoires générées au moyen d'un modèle de superpopulation. Nous illustrons la méthode à l'aide d'estimateurs par le ratio et d'estimateurs définis comme des solutions d'équations d'estimation pondérées par calage. Nous présentons aussi les résultats de simulations en vue de déterminer la performance de l'estimateur de variance proposé pour les paramètres du modèle.

Mots clés : Calage ; estimateurs par le ratio ; variance totale ; régression logistique ; équations d'estimation pondérées.

## 1. Introduction

Dans les sondages, on s'intéresse souvent à l'estimation d'un total de population finie  $Y = \sum_{k=1}^N y_k \equiv Y(y)$ , où  $N$  est la taille de la population finie. Sous un plan d'échantillonnage général avec probabilités d'inclusion positives  $\pi_k$ , un estimateur sans biais sous le plan du total  $Y$  est habituellement donné par  $\hat{Y} = \sum_{k \in s} y_k / \pi_k \equiv \sum_{k=1}^N d_k(s) y_k$ , où  $s$  est un échantillon,  $d_k(s) = a_k(s) / \pi_k$  sont les poids de sondage avec  $a_k(s) = 1$  si  $k \in s$  et  $a_k(s) = 0$  autrement. Nous utilisons la notation opérationnelle et écrivons  $Y(z) = \sum_{k=1}^N d_k(z) y_k$  de sorte que  $\hat{Y} = Y(y)$ . Donc, toutes les sommes étant considérées sur l'ensemble de la population, nous écrivons  $\sum_{k=1}^N y_k = \sum y_k$  et  $Y(z) = \sum d_k(s) z_k$  pour simplifier la notation. De nouveau, en utilisant la notation opérationnelle, nous notons un estimateur sans biais de la variance de  $Y(z)$  comme une fonction quadratique,  $9(z)$ , en termes des  $z_k$ .

Des estimateurs plus complexes d'un total  $Y$  basés sur des données auxiliaires connues de population, et des estimateurs par le ratio et par la régression, et des estimateurs de paramètres plus complexes obtenus comme

solutions d'équations d'estimation pondérées d'échantillon, tels que les estimateurs des coefficients de régression logistique en population finie, sont aussi utilisés souvent en pratique. Les estimateurs qui peuvent être exprimés sous forme d'une fonctionnalité générale  $T(\hat{M})$ , où  $\hat{M}$  désigne une mesure qui attribue le poids  $d_k(s)$  à  $y_k$ , ont également été étudiés ; par exemple,  $T(\hat{M}) = \int x d\hat{M}(x) = \sum d_k(s) y_k$  si le paramètre de population est le total  $T(M) = \int x dM(x) = \sum y_k$ , où la mesure  $M$  attribue une masse unitaire à chaque  $y_k$  (Deville 1999). L'estimation en grand échantillon de la variance d'estimateurs aussi complexes, disons  $\theta$ , a été discutée abondamment dans la littérature spécialisée. En particulier, les méthodes d'estimation de la variance de  $\theta$  par linéarisation de Taylor sont généralement applicables à tout plan d'échantillonnage qui permet d'utiliser un estimateur de variance sans biais  $9(z)$  de  $Y(z)$ . Binder (1983) a étudié les estimateurs  $\theta$  qui sont des solutions d'équations d'estimation pondérées et a appliqué la linéarisation de Taylor pour obtenir un estimateur de variance qui peut être exprimé comme  $9(z)$ , où la variable linéarisée  $z_k$  dépend de paramètres inconnus, et  $z_k$  est remplacée par un

1. Abdellatif Demnati, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa (Ontario) Canada, K1A 0T6. Courriel : Abdellatif.Demnati@statcan.gc.ca ; J.N.K. Rao, École de mathématique et de statistique, Université Carleton, Ottawa (Ontario) Canada, K1S 5B6. Courriel : jrao@math.carleton.ca.



Foster, K. (1998). Evaluating nonresponse on household surveys. *GSS Methodology Series*, 8, Office for National Statistics. Londres.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken : Wiley.

Fuller, W.A., Loughlin, M.M. et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.

Ireland, C.T., et Kuilback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.

Kalton, G., et Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.

Kalton, G., et Mahiegall, D.S. (1991). A comparison of methods for weighting adjustment for nonresponse. *Proceedings of the US Bureau of the Census 1991 Annual Research Conference*, 409-428.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 149-160.

Muenich, R., et Schultze, J. (2003). Monte Carlo simulation study of European surveys, Workpackage 3, Deliverables 3.1 and 3.2. DACSEIS project. Disponible au <http://www.unifrier.de/index.php?id=29730>.

Office for National Statistics (1998). *Labour Force Survey User Guide, Volume 1: Background and Methodology*. Londres.

Sämdal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester, Angleterre.

Stukel, D.M., Hidiroglou, M.A. et Sämdal, C.-E. (1996). Estimation de la variance des estimateurs de calage : comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 117-126.

## Remerciements

Les commentateurs de deux examinateurs nous ont aidé à améliorer considérablement le présent article. Nous remercions l'Office for National Statistics d'avoir mis à notre disposition les données de l'Enquête sur la population active, ainsi que Ralf Münich et ses collègues du projet DACSEIS (<http://www.dacseis.de/>) de nous avoir fourni la population synthétique basée sur l'Enquête sur les revenus et dépenses de l'Allemagne. La présente étude a été financée par l'Economic and Social Research Council.

## Bibliographie

Binder, D.A., et Théberge, A. (1988). Estimating the variance of raking ratio estimators. *Canadian Journal of Statistics*, 16, Supp. 47-55.

Brackstone, G.J., et Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā, Série C*, 41, 97-114.

Chang, T., et Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.

Demati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête (avec discussion). *Techniques d'enquête*, 30, 17-37.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.

Deville, J.-C., et Sämdal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-82.

Deville, J.-C., Sämdal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-20.





Tableau 6.5

Propriétés des estimateurs de variance de l'estimateur du revenu total d'après l'EDR (R = 1 000)

Méthode de pondération	Résidus pondérés par $w$	Poids utilisés pour $B$ dans le résidu <sup>1</sup>	Moyenne de l'estimateur de	Biais de l'estimateur de l'e.-t. de	REQM de l'estimateur de l'e.-t. de	Couverture <sup>2</sup> de l'intervalle de confiance (%)
------------------------	--------------------------	---	----------------------------	-------------------------------------	------------------------------------	--

Réponse complète :

Calage GREG	$d$	$d$	10 338,8	-138,5 (6,9)	259,0	93,8
$d$	$d$	$w$	10 339,2	-138,2 (6,9)	258,8	93,8
$d$	$d$	$w$	10 377,9	-99,5 (6,9)	240,0	94,1
$d$	$d$	$w$	10 376,8	-100,5 (6,9)	240,3	94,1
Calage par ratissement classique	$d$	$d$	10 338,8	-145,3 (6,9)	262,7	93,8
$d$	$d$	$w$	10 339,2	-144,9 (6,9)	262,5	93,8
$d$	$d$	$w$	10 370,0	-106,1 (6,9)	243,1	94,0
$d$	$d$	$w$	10 376,9	-107,2 (6,9)	243,5	94,0
Calage basé sur le « MV »	$d$	$d$	10 338,8	-152,7 (6,9)	266,9	93,9
$d$	$d$	$w$	10 339,2	-152,4 (6,9)	266,7	93,9
$d$	$d$	$w$	10 340,3	-151,3 (6,9)	266,1	94,0
$d$	$d$	$w$	10 378,3	-113,2 (6,9)	246,5	94,0
$w$	$w$	$w$	10 377,1	-114,4 (6,9)	247,0	94,0
$w$	$w$	$w$	10 376,7	-114,8 (6,9)	247,2	94,0

Non-réponse multiplicative :

Calage GREG	$d$	$d$	8 104,7	-5 482,1 (7,4)	5 487,1	75,8
$d$	$d$	$w$	8 105,5	-5,81,3 (7,4)	5 486,3	75,8
$d$	$d$	$w$	13 214,5	-372,3 (12,8)	549,7	94,5
$d$	$d$	$w$	13 210,9	-375,9 (12,8)	551,7	94,5
Calage par ratissement classique	$d$	$d$	8 104,7	-5 479,8 (7,4)	5 484,9	75,8
$d$	$d$	$w$	8 105,5	-5 479,1 (7,4)	5 484,1	75,8
$d$	$d$	$w$	13 214,1	-370,4 (12,8)	549,4	94,5
$d$	$d$	$w$	13 210,4	-374,2 (12,8)	551,5	94,5
Calage basé sur le « MV »	$d$	$d$	8 104,7	-5 478,1 (7,4)	5 483,1	75,8
$d$	$d$	$w$	8 105,5	-5 477,3 (7,4)	5 482,3	75,8
$d$	$d$	$w$	8 108,1	-5 474,7 (7,4)	5 479,7	75,9
$d$	$d$	$w$	13 215,2	-367,6 (12,9)	549,4	94,5
$d$	$d$	$w$	13 210,6	-372,2 (12,9)	551,6	94,5
$d$	$d$	$w$	13 208,9	-373,9 (12,9)	552,3	94,5

Non-réponse additive :

Calage GREG	$d$	$d$	8 106,3	-5 508,5 (7,4)	5 513,5	75,6
$d$	$d$	$w$	8 107,1	-5 507,7 (7,4)	5 512,7	75,6
$d$	$d$	$w$	13 207,9	-407,0 (12,8)	573,8	94,3
$d$	$d$	$w$	13 204,3	-410,5 (12,8)	575,9	94,3
Calage par ratissement classique	$d$	$d$	8 106,3	-5 506,6 (7,4)	5 511,6	75,7
$d$	$d$	$w$	8 107,1	-5 505,9 (7,4)	5 510,9	75,7
$d$	$d$	$w$	13 207,7	-405,3 (12,8)	573,6	94,1
$d$	$d$	$w$	13 203,9	-409,0 (12,8)	575,8	94,1
Calage basé sur le « MV »	$d$	$d$	8 106,3	-5 507,2 (7,4)	5 512,2	75,9
$d$	$d$	$w$	8 107,1	-5 506,4 (7,4)	5 511,4	75,9
$d$	$d$	$w$	8 109,7	-5 503,8 (7,4)	5 508,8	75,9
$d$	$d$	$w$	13 208,9	-404,6 (12,9)	574,8	94,1
$d$	$d$	$w$	13 204,2	-409,2 (12,9)	577,3	94,1
$d$	$d$	$w$	13 202,5	-411,0 (12,9)	578,1	94,1

<sup>1</sup> Voir le texte qui suit l'équation (4.8), où les choix  $df$ ,  $d$  et  $w$  correspondent à  $B$  dans (i), (ii) et (iii) respectivement.

<sup>2</sup> Pourcentage d'intervalle de confiance à 95 % de la théorie normale contenant la valeur réelle.

**Tableau 6.4** Biais relatif (%) des estimateurs de l'erreur-type des totaux de personnes chômeuses, occupées et inactives d'après l'EPA (R = 1 000)

Méthode de pondération	Résidus pondérés par $w$ ou $d^1$	Poids utilisé pour $B$ dans le résidu <sup>1</sup>	Biais relatif de l'estimateur de l'erreur-type	Chômeuses	Occupées	Inactives
------------------------	-----------------------------------	--	--	-----------	----------	-----------

Réponse complète : Calage GREG

$d$	$d$	$d$	-4,2	-3,4	0,5	0,6
$w$	$w$	$d$	-2,2	-2,2	1,9	1,7
$w$	$w$	$w$	-2,4	-2,3	0,7	0,8
$d$	$d$	$w$	-4,1	-3,2	2,1	0,8
$d$	$d$	$w$	-4,2	-3,3	0,7	0,7

Calage par ratisage classique

$d$	$d$	$w$	-2,2	-2,2	0,8	0,8
$w$	$w$	$d$	-2,2	-2,1	2,1	1,9
$w$	$w$	$w$	-2,4	-2,2	0,7	0,7

Calage basé sur le « MV »

$d$	$d$	$d$	-4,3	-3,3	0,7	0,8
$d$	$d$	$w$	-4,2	-3,3	1,1	2,3
$w$	$w$	$d$	-2,1	-2,0	1,9	1,8

Non-réponse multiplicative :

Calage GREG

$d$	$d$	$d$	-22,6	-22,3	-18,2	-18,1
$w$	$w$	$w$	-1,8	-3,3	1,8	1,5
$w$	$w$	$w$	-2,1	-3,5	1,5	1,8

Calage par ratisage classique

$d$	$d$	$d$	-22,7	-30,6	-18,4	-18,3
$d$	$d$	$w$	-22,6	-30,5	-18,3	-1,7
$w$	$w$	$d$	-1,7	-13,5	1,7	1,3

Calage basé sur le « MV »

$d$	$d$	$d$	-22,8	-22,0	-18,4	-18,3
$d$	$d$	$w$	-22,7	-21,9	-18,3	-17,9
$w$	$w$	$d$	-1,5	-2,7	1,9	1,3
$w$	$w$	$w$	-2,1	-3,1	1,3	1,1

Non-réponse additive :

Calage GREG

$d$	$d$	$d$	-22,3	-21,8	-18,5	-18,4
$w$	$w$	$w$	-1,6	-2,9	1,1	0,8
$w$	$w$	$w$	-2,0	-3,1	0,8	-18,0

Calage par ratisage classique

$d$	$d$	$d$	-22,3	-30,2	-18,0	-17,9
$d$	$d$	$w$	-22,2	-30,1	-17,9	1,8
$w$	$w$	$d$	-1,5	-13,3	1,4	1,8

Calage basé sur le « MV »

$d$	$d$	$d$	-22,4	-21,6	-18,0	-17,9
$d$	$d$	$w$	-22,3	-21,5	-17,9	-17,6
$w$	$w$	$d$	-1,3	-2,4	2,0	1,5
$w$	$w$	$w$	-1,9	-2,8	1,5	1,3

<sup>1</sup> Voir le texte qui suit l'équation (4.8), où  $df$ ,  $d$  et  $w$  correspondent à  $B$  dans (i), (ii) et (iii), respectivement.



**Tableau 6.3**  
Propriétés des estimateurs de variance pour l'estimation du nombre total de chômeurs d'après l'EPA (R = 1 000)

Méthode de	Résidus	Poids utilisés	Moyenne de	Biais de l'estimateur	REQM de	Couverture <sup>2</sup> de
pondération	pondérés par w	pour B dans le résidu <sup>1</sup>	l'estimateur de l'erreur-type	de l'e.-t. de simulation)	l'estimateur de l'e.-t.	l'intervalle de confiance (%)

Réponse complète :

Calage GREG

d 433,9 -18,8 (0,9) 33,4 33,3 93,5

d 434,3 -18,5 (0,9) 33,3 31,9 93,8

w 442,8 -10,0 (1,0) 32,0 32,0 93,8

w 441,9 -10,8 (1,0) 32,0 32,0 93,7

d 433,9 -18,8 (0,9) 33,4 33,3 93,5

d 434,2 -18,5 (0,9) 33,3 33,3 93,5

w 443,0 -9,8 (1,0) 32,0 32,0 93,8

w 442,0 -10,7 (1,0) 32,0 32,0 93,8

d 433,9 -19,4 (0,9) 33,7 33,7 93,5

d 434,3 -19,1 (0,9) 33,6 33,6 93,5

d 435,4 -17,9 (0,9) 33,0 33,0 93,5

d 443,7 -9,6 (1,0) 32,5 32,5 93,7

w 442,3 -11,1 (1,0) 32,4 32,4 93,7

w 441,6 -11,8 (1,0) 32,3 32,3 93,7

Non-réponse multiplicative :

Calage GREG

d 385,7 -112,6 (0,9) 116,0 115,5 85,8

d 386,1 -112,1 (0,9) 115,5 115,5 85,8

w 489,5 -8,8 (1,2) 39,2 39,2 94,2

w 487,8 -10,4 (1,2) 39,2 39,2 94,2

d 385,7 -113,1 (0,9) 116,5 116,5 85,7

w 386,1 -112,7 (0,9) 116,1 116,1 85,7

w 490,3 -8,5 (1,2) 39,6 39,6 94,3

w 488,4 -10,4 (1,2) 39,5 39,5 94,1

d 385,7 -113,7 (0,9) 117,1 117,1 85,4

d 386,2 -113,2 (0,9) 116,6 116,6 85,6

d 387,8 -111,6 (0,9) 115,0 115,0 85,8

d 491,9 -7,5 (1,3) 40,4 40,4 94,2

w 488,9 -10,5 (1,2) 39,9 39,9 94,0

w 487,5 -11,9 (1,2) 39,8 39,8 94,0

Non-réponse additive :

Calage GREG

d 386,5 -110,9 (0,9) 114,4 113,9 86,0

w 387,0 -110,5 (0,9) 113,9 113,9 86,0

d 489,3 -8,2 (1,2) 39,0 39,0 94,6

w 487,6 -9,8 (1,2) 39,0 39,0 94,6

d 386,5 -111,0 (0,9) 114,4 114,4 85,8

w 387,0 -110,6 (0,9) 114,0 114,0 85,8

w 490,1 -7,4 (1,2) 39,2 39,2 94,7

w 488,1 -9,4 (1,2) 39,1 39,1 94,6

d 386,5 -111,6 (0,9) 115,0 115,0 85,6

d 387,0 -111,1 (0,9) 114,6 114,6 85,6

d 388,6 -109,5 (0,9) 113,0 113,0 85,9

d 491,6 -6,5 (1,3) 40,0 40,0 94,7

w 488,6 -9,5 (1,2) 39,5 39,5 94,6

w 487,3 -10,8 (1,2) 39,4 39,4 94,6

Calage basé sur le « MV »

<sup>1</sup> Voir le texte qui suit l'équation (4,8), où les choix  $df$ ,  $d$  et  $w$  correspondent à  $B$  dans (i), (ii) et (iii) respectivement.  
<sup>2</sup> Pourcentage d'intervalle de confiance à 95 % de la théorie normale contenant la valeur réelle.

### 6.2 Propriétés des estimateurs de variance

Les propriétés des divers estimateurs des variances des estimateurs ponctuels du nombre total de chômeurs d'après l'EPA sont présentées au tableau 6.3 (dans ce tableau, l'« estimation de l'erreur-type » désigne la racine carrée de l'estimation de la variance). Nous faisons plusieurs observations :

- la pondération des résidus par  $w_i$  plutôt que par  $d_i$  réduit le biais et la racine carrée de l'erreur quadratique moyenne de l'estimateur de l'erreur-type. Le biais dû à l'utilisation des résidus pondérés par  $d_i$  est particulièrement important en cas de non-réponse (comme l'a fait remarquer Fuller 2002), mais nous constatons des réductions non négligeables du biais même en cas de réponse complète ;
- le choix du poids utilisé dans  $B$  pour le calcul des résidus semble avoir peu d'effet ;
- pour un modèle de non-réponse et un choix de pondération des résidus donnés, les résultats produits par les divers choix de l'estimateur ponctuel diffèrent peu.

Au tableau 6.4, les résultats du tableau 6.3 sont étendus afin de prendre en considération le biais relatif des estimateurs de l'erreur-type, plutôt que leur biais absolu et pour considérer deux paramètres supplémentaires, à savoir les nombres totaux de chômeurs et de personnes inactives. Nous voyons de nouveau que le biais relatif dû à l'utilisation des résidus pondérés par  $d_i$  peut être important en présence de non-réponse, supérieur à 20 % dans plusieurs cas, et qu'il est réduit en utilisant les résidus pondérés par  $w_i$ . Encore une fois, nous observons peu de changements du biais relatif en pourcentage des estimateurs de l'erreur-type quand nous utilisons divers choix de pondération dans le calcul de  $B$  pour les résidus.

Les résultats correspondants pour les données de l'ERD lorsque l'on estime le revenu total sont présentées au tableau 6.5. De nouveaux, les résultats sont d'allure généralement similaire à ceux obtenus pour les données de l'EPA au tableau 6.3. En cas de réponse complète, l'utilisation des résidus pondérés par  $w_i$  plutôt que par  $d_i$  produit une légère amélioration du biais et de la RQM des estimateurs de l'erreur-type. Pour les cas de non-réponse, les améliorations sont considérables. Nous observons peu de changements dans les estimateurs de l'erreur-type quand nous modifions le choix de la pondération utilisée pour estimer les coefficients de régression. Au tableau 6.6, les résultats

### 7. Conclusion

du tableau 6.5 sont étendus afin de prendre en considération le biais relatif des estimateurs de l'erreur-type plutôt que leur biais absolu et de considérer un paramètre supplémentaire, à savoir les dépenses totales par trimestre. De nouveau, nous voyons que le biais relatif résultant de l'utilisation des résidus pondérés par  $d_i$  peut être important en présence de non-réponse, supérieur à 35 % dans tous les cas, et qu'il est réduit si l'on utilise les résidus pondérés par  $w_i$  pour lesquels le biais relatif n'est jamais supérieur à environ 3 %.

L'étude par simulation a révélé peu de différences entre le biais ou les propriétés de variance des trois estimateurs par calage considérés, à savoir l'estimateur GREG, l'estimateur par calage classique (ratisage croisé) et l'estimateur par calage basé sur le maximum de vraisemblance. Nous avons observé certains petits écarts dans la distribution des poids extrêmes, l'estimateur par calage fondé sur le maximum de vraisemblance produisant le plus grand nombre de poids très grands et l'estimateur GREG étant le seul produisant quelques poids négatifs.

En ce qui concerne les estimateurs de variance, la principale observation est le contraste entre l'approche consistant à pondérer les résidus par les poids de sondage et celle consistant à les pondérer par les poids calés. Nous avons constaté que le second estimateur de variance possédait systématiquement un biais plus petit et que cet effet était très marqué en présence de non-réponse, situation dans laquelle le premier estimateur pouvait être gravement biaisé. Le biais du second estimateur était généralement petit et le niveau de couverture des intervalles de confiance associés était généralement proche de la couverture nominale.

Nous avons considéré d'autres moyens de pondérer les observations pour construire les coefficients de régression lorsque l'on calcule les résidus de l'estimateur de variance par linéarisation, mais les effets observés étaient faibles et nous n'avons recueilli aucune preuve que ce choix est important en pratique. En général, les constatations concernant les variables catégoriques dans l'Enquête sur la population active du Royaume-Uni étaient remarquablement comparables à celles pour les variables continues de l'Enquête sur les revenus et dépenses de l'Allemagne.

méthode de calage GREG produit certains points négatifs, tandis que les deux méthodes de ratisage permettent d'éviter ce problème, comme prévu. Un plus grand nombre de poids très grands sont toutefois observés pour l'estimateur par calage basé sur le « maximum de vraisemblance ».

Les résultats correspondants pour les données de l'ERD sont présentées au tableau 6.2. La tendance des résultats est généralement similaire, quoiqu'il n'y ait pas d'évidence de présence d'un biais de non-réponse significatif (c'est-à-dire que le biais observé peut être expliqué par des variations de simulation). Les erreurs-types et les racines carrées des erreurs quadratiques moyennes demeurent également presque constantes d'une méthode de pondération à l'autre pour un modèle de non-réponse donné.

Le tableau 2 donne des preuves d'un biais de non-réponse, qui est d'ordre similaire pour chacune des méthodes de calage. Nous ne constatons pas que ce biais est moindre quand l'estimateur concorde avec le modèle de non-réponse (c'est-à-dire l'estimateur GREG pour la réponse additive et l'estimateur par calage (ratisage croisé) pour la réponse multiplicative) comme nous aurions pu nous y attendre. Cela pourrait tenir au fait que les covariables utilisées dans les modèles de non-réponse (par exemple, la variable d'âge égal ou supérieur à 35 ans) ne sont pas toutes incluses dans les variables de calage. Néanmoins, le biais de non-réponse est faible en ce sens que la racine carrée de l'erreur quadratique moyenne est très semblable à l'erreur-type dans chaque cas. En présence de non-réponse, la

Tableau 6.1

Propriétés de simulation des estimateurs ponctuels du total de chômeurs en utilisant les données de l'EPA avec R = 1 000

Modèle de non-réponse/estimateur ponctuel						
Biais	Erreur-type	Racine carrée de l'erreur	Nombre de poids négatifs <sup>1</sup>	Nombre de poids très grands <sup>1,2</sup>		
(erreur-type de simulation)		quadratique moyenne				
Réponse complète :						
Calage GREG	7,6 (14,3)	452,8	452,8	0	0	0
Calage par ratisage classique	8,3 (14,3)	452,8	452,9	0	0	0
Calage basé sur le « MV »	9,0 (14,3)	453,3	453,4	0	1	1
Non-réponse multiplicative :						
Calage GREG	-45,6 (15,8)	498,3	500,3	4	1	1
Calage par ratisage classique	-42,1 (15,8)	498,8	500,6	0	2	2
Calage basé sur le « MV »	-39,7 (15,8)	499,4	501,0	0	7	7
Non-réponse additive :						
Calage GREG	-37,3 (15,7)	497,4	498,8	5	1	1
Calage par ratisage classique	-34,7 (15,7)	497,5	498,7	0	3	3
Calage basé sur le « MV »	-32,4 (15,8)	498,1	499,1	0	7	7

<sup>1</sup>Nombre de ces poids sur l'ensemble des unités échantillonnées sur l'ensemble des 1 000 échantillons.

<sup>2</sup>Nombre de poids égaux à plus de dix fois le poids de sondage correspondant.

Tableau 6.2

Propriétés de simulation des estimateurs ponctuels du revenu total en utilisant les données de l'ERD avec R = 1 000

Modèle de non-réponse/estimateur ponctuel						
Biais	Erreur-type	Racine carrée de l'erreur	Nombre de poids négatifs	Nombre de poids très grands		
(erreur-type de simulation)		quadratique moyenne				
Réponse complète :						
Calage GREG	-172,2 (331,3)	10 477,3	10 478,7	0	0	0
Calage par ratisage classique	-170,6 (331,5)	10 484,1	10 485,8	0	0	0
Calage basé sur le « MV »	-169,8 (331,8)	10 491,5	10 492,9	0	0	0
Non-réponse multiplicative :						
Calage GREG	-495,7 (429,7)	13 586,8	13 595,8	0	0	0
Calage par ratisage classique	-493,8 (429,6)	13 584,6	13 593,5	0	0	0
Calage basé sur le « MV »	-463,5 (429,5)	13 582,8	13 590,7	0	0	0
Non-réponse additive :						
Calage GREG	-473,2 (430,5)	13 614,8	13 623,0	0	0	0
Calage par ratisage classique	-469,4 (430,5)	13 612,9	13 621,0	0	0	0
Calage basé sur le « MV »	-439,5 (430,5)	13 613,5	13 620,6	0	0	0



- une classification croisée de la région (Nord de l'Angleterre ; Londres et Sud-Est ; Midlands et East Anglia ; Ecosse) selon le sexe et l'âge par tranche de 15 ans (16 à 29 ans, 30 à 44 ans, 45 à 59 ans, 60 à 75 ans et plus) avec 40 catégories.

### 5.2 Étude fondée sur l'Enquête sur les revenus et dépenses de l'Allemagne

Notre deuxième étude est basée sur l'édition de 1998 de l'Enquête sur les revenus et dépenses (ERD) de l'Allemagne, qui est une enquête-ménage nationale réalisée tous les cinq ans par le Bureau fédéral de la statistique pour fournir des renseignements sur la situation économique et sociale des ménages, surtout en ce qui concerne la distribution des revenus et des dépenses (Muennich et Schultze 2003). Nous avons utilisé les données provenant d'une population synthétique de 64 326 ménages, créée pour représenter 20 % des ménages de la région de Brême à l'exclusion de ceux dont le revenu mensuel net du ménage était égal ou supérieur à 35 000 DM (DM désigne le mark allemand). Un plan d'échantillonnage par quota a été utilisé pour cette enquête et nous n'avons pas essayé de le calquer. À la place, nous avons utilisé dans notre simulation un échantillonnage aléatoire simple avec un modèle de non-réponse. Nous avons tiré des échantillons aléatoires simples répétés de 1 340 ménages à partir de la population artificielle, ce qui représente une fraction d'échantillonnage d'environ 1/48. Nous avons construit les modèles de non-réponse en utilisant les résultats d'études portant sur des enquêtes similaires réalisées en Grande-Bretagne, à savoir l'Enquête sur les dépenses des familles et l'Enquête nationale sur les aliments (Foster 1998). Pour chaque échantillon sélectionné, nous avons déterminé le sous-ensemble de ménages répondants à l'aide des modèles de non-réponse suivants :

*Modèle multiplicatif :*

$$q_i^{-1} = 1,44 \times 1,09 \text{ (si travailleur autonome)} \times 1,03 \text{ (si chômeur)} \times 0,97 \text{ (si travailleur)} \times 1,16 \text{ (si aucun enfant dans le ménage)}.$$

*Modèle additif :*

$$q_i^{-1} = 1,44 + 0,13 \text{ (si travailleur autonome)} + 0,04 \text{ (si chômeur)} + 0,04 \text{ (si travailleur)} + 0,23 \text{ (si aucune enfant dans le ménage)}.$$

Les paramètres d'intérêt sont le revenu net total du ménage par trimestre et les dépenses totales du ménage par trimestre, calculés d'après les données sur la population finie artificielle. Comme pour l'étude fondée sur l'EPA, nous avons attribué un poids à chaque ménage échantillonné. Dans l'ERD réelle, les poids sont construits essentiellement selon la

méthode de calage basée sur le maximum de vraisemblance en ajustant simultanément les données d'échantillon aux distributions marginales de plusieurs caractéristiques, telles que le type de ménage, le statut socioéconomique de la personne de référence, la catégorie de revenu net du ménage et la région (land). Nous essayons d'imiter ce redressement dans la mesure du possible dans notre étude. Toutefois, comme dans le cas de l'EPA, en raison du problème que posent les strates contenant un petit nombre de ménages, nous simplifions les variables de calage de l'ERD pour obtenir les trois facteurs catégoriques suivants :

- type de ménage avec sept catégories  
— mère ou père seul plus un enfant,  
— mère ou père seul plus deux enfants ou plus,  
— couple avec un enfant – conjoint travailleur,  
— couple avec un enfant – conjoint chômeur,  
— couple avec deux enfants ou plus – conjoint travailleur,  
— couple avec deux enfants ou plus – conjoint chômeur,  
— autre ;
- statut social de la personne de référence avec cinq catégories  
— travailleur autonome,  
— fonctionnaire ou militaire,  
— employé,  
— ouvrier,  
— chômeur, pensionné, étudiant ou autre ;
- revenu net du ménage par trimestre avec trois catégories  
— 0 à 5 000 DM,  
— 5 000 à 7 000 DM,  
— 7 000 à 35 000 DM.

## 6. Résultats

### 6.1 Propriétés des estimateurs ponctuels

Le tableau 6.1 donne les propriétés des estimateurs ponctuels du nombre total de chômeurs dans l'étude de l'EPA pour diverses méthodes de calage et diverses hypothèses au sujet de la non-réponse. Les propriétés sont évaluées selon les pratiques habituelles dans les études par simulation. Par exemple, au tableau 6.1, le biais est calculé d'après  $B(\hat{T}_y) = E(\hat{T}_y) - T_y$ , où  $E(\hat{T}_y) = 1/R \sum_{r=1}^R \hat{T}_{y,r}$ ,  $\hat{T}_{y,r}$  est la valeur de  $\hat{T}_y$  pour l'échantillon  $r$  et  $R$  est le nombre d'échantillons simulés. Nous constatons, en examinant ce tableau, que l'erreur-type demeure presque constante pour les diverses méthodes de calage pour un modèle de non-réponse donné. La non-réponse accroît l'erreur-type dans tous les estimateurs comme il fallait s'y attendre (puisque la taille d'échantillon

$T_y$  par estimation de la variance de l'estimateur linéaire  $\sum_i d_i z_i$ , en traitant  $d_i$  et  $z_i$  comme étant fixes. Dans le cas d'un plan d'échantillonnage à plusieurs degrés stratifié, en supposant que le tirage des unités primaires d'échantillonnage (UPE) dans les strates se fait « avec remise », un estimateur standard de la variance (par exemple, Stukel et coll. 1996) est donné par :

$$V(T_y) = \sum_h \frac{n_h}{n} - 1 \sum_h (z_h - \bar{z})^2 \quad (4.9)$$

où  $z_h = \sum_k d_{hjk} z_{hjk}$ ,  $\bar{z}_h = \sum_j z_{hj}/n_h$  et  $z_{hjk}$  est la valeur de la variable définie en (4.8) pour le  $k^e$  individu dans la  $j^e$  UPE sélectionnée dans la strate  $h$ . Cet estimateur demeure approprié en présence de non-réponse si la réponse individuelle dans chaque UPE est indépendante de la réponse dans toutes les autres UPE et qu'au moins un individu est observé dans chaque UPE sélectionnée (Fuller et coll. 1994, page 78).

## 5. Études par simulation

Afin de comparer la performance des estimateurs pondérés à celle des estimateur de variance correspondants, nous avons effectué deux études par simulation en construisant des populations artificielles en nous servant des données de l'Enquête sur la population active (EPA) du Royaume-Uni et de l'Enquête sur les revenus et les dépenses (ERD) de l'Allemagne. Dans chaque cas, nous avons généré  $R = 1\,000$  échantillons à partir de ces populations en procédant d'abord à l'échantillonnage de manière à calquer le plan de sondage réel moyennant certaines simplifications, puis en éliminant les cas de non-réponse conformément à deux modèles de non-réponse. Le premier modèle suppose une non-réponse de forme multiplicative qui, d'après la condition C de la section 3, pourrait donner lieu à un biais plus faible pour la méthode de calage par le raking ratio. Le deuxième modèle s'appuie sur l'hypothèse d'une non-réponse additive pour pourrait donner lieu à un biais plus faible pour l'estimateur GREG.

Pour chacun des  $R$  échantillons, nous avons calculé les estimations ponctuelles des paramètres en nous servant des diverses méthodes de calage généralisé présentées à la section 2 et estimé les variances en utilisant les diverses méthodes par linéarisation présentées à la section 4. Ensuite, nous avons résumé les propriétés des estimateurs.

### 5.1 Étude fondée sur l'Enquête sur la population active du Royaume-Uni

La première étude s'appuyait sur des données provenant du trimestre de mars à mai 1998 de l'EPA du Royaume-Uni, qui est une enquête auprès de la population à domicile du Royaume-Uni, conçue pour fournir des renseignements

sur le marché britannique du travail et réalisée par l'Office for National Statistics (ONS). Nous avons traité l'échantillon d'environ 58 000 ménages comme une population artificielle. Nous avons tiré des échantillons répétés de cette population de façon à imiter le plan de sondage utilisé pour l'EPA (ONS 1998, section 3). Chaque échantillon comprend 1 211 ménages sélectionnés par échantillonnage aléatoire stratifié avec répartition proportionnelle entre 19 strates définies selon la région de résidence. Ces régions ont été définies de façon qu'elles correspondent aux secteurs affectés aux intervieweurs, qui définissaient les strates dans l'EPA. Dans le cadre de cette enquête, toutes les personnes faisant partie d'un ménage échantillonné sont interviewées dans la mesure du possible. Dans notre étude par simulation, nous avons retenu tous les répondants compris dans un ménage échantillonné, sauf ceux de moins de 16 ans, qui sont sans pertinence pour les estimations d'intérêt.

Pour déterminer si les personnes échantillonnées avaient répondu, nous nous sommes servis des deux modèles de non-réponse qui suivent, basés sur les résultats d'une étude de Foster (1998).

*Modèle de non-réponse multiplicatif :*

$$q_i^{-1} = 1,15 \times 1,17 \text{ (si Londres)} \\ \times 1,13 \text{ (si moins de 35 ans)} \\ \times 1,1 \text{ (si sexe féminin)}$$

*Modèle de non-réponse additif :*

$$q_i^{-1} = 1,15 + 0,20 \text{ (si Londres)} \\ + 0,15 \text{ (si moins de 35 ans)} \\ + 0,10 \text{ (si sexe féminin)}$$

Trois paramètres d'intérêt sont définis pour la population artificielle : les nombres totaux de personnes en chômage, répondants individuels en nous servant de totaux de calage qui correspondaient aux dénombrements de population dans les catégories des trois variables auxiliaires catégoriques et des poids initiaux d'Horvitz-Thompson  $d_i$ , comme à la section 2. Ce choix des variables auxiliaires avait pour but de calquer celles utilisées dans l'EPA. Cependant, étant donné l'échelle réduite de notre population artificielle et les nombres conséquemment plus petits de personnes dans les strates, nous avons simplifié les variables de calage de l'EPA pour obtenir les trois facteurs catégoriques suivants, qui définissent 83 totaux de contrôle :

- région de résidence avec 23 catégories ;
- une classification croisée du sexe selon dix groupes d'âge (consistant en années uniques pour les personnes de 16 à 24 ans et en un groupe d'âge distinct pour les 25 ans et plus) avec 20 catégories ;



et nous supposons que  $B$  converge vers une matrice limite finie  $\beta$ . Une autre dérivation de cette expression est donnée par Dermani et Rao (2004, section 3.4). Pour l'estimation de la variance par linéarisation,  $\hat{T}_y$  est traité comme étant l'estimateur linéaire  $\sum_s d_i z_i$ , où

$$z_i = F_i'(y_i - \beta x_i) \quad (4.8)$$

est traité comme une variable fixe. Un certain nombre de choix de  $F_i$  et  $\beta$  ont été discutés dans la littérature. En ce qui concerne  $F_i$ , le choix naturel impliqué par l'argument qui précède est  $F_i = F(x_i' \lambda)$ . Toutefois, un choix plus simple consisterait à prendre  $F_i = 1$ . Deville et Särndal (1992) soulignent que, dans leur théorie classique avec  $\lambda = 0$ , ces choix sont asymptotiquement équivalents, mais expriment une préférence pour  $F_i = F(x_i' \lambda)$ . Sous nos conditions selon lesquelles il existe une non-réponse et où l'égalité  $\lambda = 0$  n'est pas nécessairement vérifiée, le deuxième choix semble préférable et c'est ce qui est souligné par Fuller (2002, page 15). Notons que ces deux choix impliquent que  $\sum_s d_i z_i$  prend la forme  $\sum w_i (y_i - \beta x_i)$  quand  $F_i = F(x_i' \lambda)$  ou  $\sum d_i (y_i - \beta x_i)$  quand  $F_i = 1$ . Nous désignerons donc ces choix comme étant les *résidus pondérés* par  $w_i$  ou les *résidus pondérés par  $d_i$* .

En ce qui concerne  $\beta$ , il découle de notre argument concernant les choix de  $F_i$  que  $f_i$  dans (4.2) devrait être remplacée par  $f_i = f(x_i' \lambda)$ , ce qui donne :

$$i) \quad B = [\sum_i d_i f_i y_i x_i'] [\sum_i d_i f_i x_i x_i']^{-1}, \text{ comme l'ont également proposé Dermani et Rao (2004).}$$

Les autres choix sont :

$$ii) \quad B = \hat{\beta}_s, \text{ comme dans (2.4), comme l'ont proposé Deville et coll. (1993).}$$

$$iii) \quad B = [\sum_i w_i y_i x_i'] [\sum_i w_i x_i x_i']^{-1}, \text{ comme l'ont proposé Deville et Särndal (1992, équation 3.4), ce qui pourrait être plus facile à calculer que } \hat{\beta}_s \text{ pour les utilisateurs des fichiers de données d'enquête qui contiennent les poids } w_i, \text{ mais non les poids } d_i.$$

La mesure dans laquelle ces choix diffèrent dépend du choix de la fonction  $G(\cdot)$ . Dans le cas linéaire,  $f(u) = 1$ , de sorte que les estimateurs donnés en (i) et (ii) sont identiques. Dans le cas du redressement par calage classique,  $f(u) = F(u) = \exp(u)$  de sorte que  $f_i = F_i$  et  $d_i f_i = w_i$  et les estimateurs (i) et (iii) sont identiques. Pour l'estimateur par calage basé sur le « maximum de vraisemblance », nous avons  $F(u) = (1 - u)^{-1}$  et  $f(u) = (1 - u)^{-2}$ , de sorte que  $d_i f_i = w_i^2 / d_i$  et les trois estimateurs de variance sont tous distincts. Après avoir déterminé la forme de  $z_i$  dans (4.8), nous obtenons l'estimateur de la variance par linéarisation pour

$$N^{-1} \sum d_i F(x_i' \lambda) x_i =$$

$$N^{-1} \sum d_i F_i x_i + N^{-1} \sum d_i f_i x_i x_i' (\hat{\lambda} - \lambda) + o_p(n^{0.5}), \quad (4.1)$$

où  $f_i = f(x_i' \lambda)$ . Alors, en supposant que  $\lim_{N \rightarrow \infty} N^{-1} \sum d_i q_i f_i x_i x_i'$  est non singulière et en utilisant (2.3), nous obtenons

$$\hat{\lambda} - \lambda = \left[ \sum d_i f_i x_i x_i' \right]^{-1} \left[ T_x - \sum d_i F_i x_i \right] + o_p(n^{0.5}). \quad (4.2)$$

Voir Fuller (2009, preuve du théorème 1.3.9) pour une description formelle de la façon dont (4.1) et (4.2) peuvent être dérivées et des conditions de régularité sous-jacentes. Notons que, pour s'assurer que  $\lim_{N \rightarrow \infty} N^{-1} \sum d_i q_i f_i x_i x_i'$  est non singulière, il pourrait être nécessaire d'éliminer de  $x_i$  les variables redondantes et peut-être (comme dans Deville et Särndal 1992) de modifier l'estimateur pour les échantillons dont la probabilité est faible qui rendent cette matrice singulière.

Un argument similaire comportant le développement en série de Taylor de  $w_i$  dans (2.2) autour de  $\lambda$  donne :

$$w_i = d_i [F_i + f_i x_i' (\hat{\lambda} - \lambda)] + o_p(N^{-1.5}). \quad (4.3)$$

Alors, en supposant que les moments de population nécessaires existent afin que le terme résiduel dans (4.3) soit vérifié uniformément sur les  $i$  (Fuller 2009, Corollaire 2.7.1.1.), nous avons

$$\hat{T}_y \equiv \sum w_i y_i$$

$$= \sum d_i [F_i + f_i x_i' (\hat{\lambda} - \lambda)] y_i + o_p(Nn^{-0.5}) \quad (4.4)$$

et, donc, découlant de (4.2) et (4.4) :

$$\hat{T}_y = \sum d_i F_i y_i + B \left[ T_x - \sum d_i F_i x_i \right] + o_p(Nn^{-0.5}), \quad (4.5)$$

où

$$B = \left[ \sum d_i f_i y_i x_i' \right] \left[ \sum d_i f_i x_i x_i' \right]^{-1}. \quad (4.6)$$

Notons que  $F_i = f_i = 1$  sous les hypothèses de Deville et Särndal (1992) (puisque, dans ce cas,  $\lambda = 0$  et il découle des hypothèses au sujet de  $G(\cdot)$  que  $F(0) = f(0) = 1$ ). Donc, sous ces hypothèses, l'expression (4.5) correspond au résultat 5 de Deville et Särndal (1992), c'est-à-dire que l'estimateur par calage généralisé est asymptotiquement équivalent à l'estimateur GREG. Par conséquent, la variance asymptotique de  $\hat{T}_y$  est la même que celle de  $\sum_s d_i z_i$ , où  $z_i$  est la variable linéarisée :

$$z_i = F_i'(y_i - \beta x_i), \quad (4.7)$$



interprétée comme une quantité proportionnelle à moins une log-vraisemblance dans le cas de l'échantillonnage aléatoire simple avec remise (Brackstone et Rao 1979 ; Fuller 2002).

### 3. Cadre asymptotique et biais de non-réponse

Nous examinons maintenant les propriétés asymptotiques de  $\hat{T}_y$  en regard du plan d'échantillonnage ainsi que du mécanisme de non-réponse. Nous supposons que ce dernier est tel que chaque unité de la population répond, si elle est échantillonnée, avec la probabilité  $q_i$ , où cette probabilité est indépendante du choix de l'échantillon et où les diverses unités répondent de manière indépendante. Nous considérons un cadre asymptotique défini en fonction de suites de populations finies et de plans d'échantillonnage probabilités et de mécanismes de réponse connexes (Fuller 2009, section 1.3), avec les termes d'ordre de grandeur exprimés en fonction de  $n = \sum_{i \in U} \pi_i q_i$ , le nombre prévu d'unités répondantes, et de  $N$ , la taille de la population. Nous supposons qu'il existe des constantes positives  $K_1, K_2$  et  $K_3$  telles que  $K_1 < nN^{-1}d_i < K_2$  et  $K_3 < q_i$  pour tout  $i$ .

Nous supposons que les estimateurs d'Horvitz-Thompson des moyennes convergent vers les moyennes de la population finie correspondantes et que le théorème de la limite centrale est vérifié (tel qu'exprimé formellement dans les conditions du théorème 1.3.9 de Fuller 2009). En particulier, nous supposons que les suites et la fonction  $F(\cdot)$  sont telles qu'il existe une solution unique  $\lambda$  de

$$(3.1) \quad \sum_{i \in U} q_i F(x'_i \lambda) x_i = T_y,$$

avec

$$(3.2) \quad \lambda = \lambda + O_p(n^{-0.5}),$$

et que

$$(3.3) \quad \hat{T}_y = \sum_{i \in U} q_i F(x'_i \lambda) y_i + O_p(Nn^{-0.5}).$$

Si la condition C est vérifiée,  $\alpha$  est une solution de (3.1) et donc  $\lambda = \alpha$ . Il découle de (3.3) que  $\hat{T}_y$  converge vers  $T_y$  pour tout choix de la variable  $y$  si cette condition est vérifiée. Donc, nous pouvons considérer la condition C

*Condition C* : il existe un vecteur  $\alpha$  tel que  $F(x'_i \alpha) = q_i^{-1}$ .

L'une de nos hypothèses clés sera :

vecteur nul.

Fuller 2002, page 15) et nous n'exigeons pas que  $\lambda$  soit le la non-réponse, cette dernière est souvent peu plausible (voir  $N^{-1}(T_y^{xx} - T_y)$   $\rightarrow 0$  en probabilité. Toutefois, dans le cas de plan d'échantillonnage et comprennent la contrainte que thèses ne s'appliquent qu'à la distribution induite par le certaines hypothèses (leur résultat 2). Toutefois, leurs hypothèses et Särndal (1992) montrent que  $\lambda = 0$  sous

comme une condition suffisante de l'absence de biais de non-réponse (asymptotique). Cette propriété de la condition C a été discutée par Fuller, Loughlin et Baker (1994), Fuller (2009, page 284) ainsi que Särndal et Lundström (2005, proposition 9.2) pour le cas où  $F$  est linéaire. Fuller (2002, page 15), Kott (2006), ainsi que Chang et Kott (2008) considèrent aussi l'estimation des probabilités de réponse en utilisant des modèles généraux de la forme  $q_i^{-1} = F(x'_i \alpha)$ . Afin d'illustrer ce qui pourrait arriver si la condition C n'était pas vérifiée, supposons que  $x_i$  est simplement une grandeur scalaire avec  $x_i \equiv 1$ . Alors, la solution unique de (3.1) est  $\lambda = g(N/\sum_{i \in U} q_i)$  et  $p \lim(\hat{T}_y) = N(\sum_{i \in U} q_i y_i) / (\sum_{i \in U} q_i)$ . D'où, le biais asymptotique de non-réponse qui ne disparaîtra que pour les variables étudiées qui sont « non corrélées » aux probabilités de réponse  $q_i$ .

### 4. Estimation de la variance par linéarisation

Nous nous penchons maintenant sur la variance asymptotique de  $\hat{T}_y$  et sur son estimation. Comme à la section précédente, nous définissons la variance par rapport à la loi conjointe induite par l'échantillonnage ainsi que la non-réponse. Commençons par noter qu'en général (et en particulier pour  $G_M(\cdot)$  et  $G_{ML}(\cdot)$ ), une itération est nécessaire pour résoudre les équations de calage. On trouve dans la littérature (voir Deville et coll. 1993) des travaux visant à estimer la variance de  $\hat{T}_y$  après un nombre fini d'itérations. Nous suivons plutôt l'approche de Deville et coll. (1993) et, par exemple, de Binder et Thèberge (1988) en approximant la variance de  $\hat{T}_y$  par la variance de l'estimateur « convergé », c'est-à-dire l'estimateur hypothétique issu d'un nombre infini d'itérations, représenté par  $\text{var}(\sum_{i \in U} w_i y_i)$ , où les  $w_i$  sont les poids « convergés » qui sont les solutions du problème d'optimisation sous contrainte de la section 2.

Un estimateur de variance par linéarisation s'obtient en approximant  $\text{var}(\sum_{i \in U} w_i y_i)$  par  $\text{var}(\sum_{i \in U} d_i z_i)$  pour une « variable linéarisée »  $z_i$  (Deville 1999). Nous cherchons maintenant à construire cette variable en utilisant un argument sous grand échantillon. Nous obtenons d'abord une expression de  $\lambda$ . Un développement en série de Taylor du premier membre des équations de calage (2.3) donne

$$\sum_{i \in U} d_i F(x'_i \lambda) x_i = \sum_{i \in U} d_i F_i x_i + \sum_{i \in U} d_i f(x'_i \lambda^*) x_i x_i' (\lambda - \lambda_i),$$

où  $F_i = F(x'_i \lambda_i)$ ,  $\lambda^*$  est compris entre  $\lambda_i$  et  $\lambda$ , et il est supposé que  $f(u) = dF(u)/du$  existe. En supposant que  $f(\cdot)$  est continue et que  $\lim_{N \rightarrow \infty} N^{-1} \sum_{i \in U} q_i f_i x_i x_i'$  existe, et en utilisant (3.2), nous avons

d'estimation (par exemple, Kalton et Maligalig 1991 ; Kalton et Flores-Cervantes 2003), et le choix de l'estimateur de variance pourrait importer davantage en présence de non-réponse (par exemple, Fuller 2002, section 8).

La plan de l'article est le suivant : les estimateurs par calage généralisé sont définis à la section 2 et, après la présentation d'un cadre asymptotique, le biais de ces estimateurs est examiné à la section 3. Les estimateurs de variance par linéarisation sont définis à la section 4. L'étude par simulation est présentée à la section 5, les résultats sont discutés à la section 6 et certaines conclusions sont énoncées à la section 7.

## 2. Estimation par calage généralisé

Nous considérons la classe des estimateurs pondérés d'un total de population  $T_y = \sum_U y_i$ , qui peut être exprimé sous la forme  $\bar{Y}_y = \sum_s w_i y_i$ , où  $y_i$  est la valeur d'une variable étudiée pour une unité  $i$  dans un échantillon  $s$  tiré d'une population  $U$  et  $w_i$  est le poids de sondage qui peut dépendre de l'échantillon, mais non du choix de la variable étudiée. Nous supposons ici que l'échantillon  $s$  est constitué de l'ensemble restant de répondants après l'échantillonnage et l'éventuelle non-réponse totale. Le calage généralisé est une forme d'estimation pondérée qui peut être employée quand l'information auxiliaire au niveau de la population est disponible sous la forme d'un vecteur  $T_x = \sum_U x_i$  des totaux de population des valeurs  $x_i$  d'un vecteur de variables auxiliaires, où la valeur  $x_i$  est connue pour toutes les unités présentes dans  $s$ . À l'instar de Deville et Särndal (1992), nous disons que les poids  $w_i$  sont *calés* s'ils satisfont aux *équations de calage*  $\sum_s w_i x_i = T_x$ . Le vecteur  $T_x$  est appelé *vecteur des totaux de calage*. La classe des poids de calage généralisés  $w_i$  est obtenue en minimisant la fonction

$$\sum_s d_i G(w_i/d_i), \quad (2.1)$$

sous la contrainte que les poids  $w_i$  soient calés, où  $G(\cdot)$  est une fonction objectif spécifiée qui satisfait à certains critères (voir Deville et coll. 1993) et  $d_i$  est un poids initial. Nous le prendrons ici égal au poids de sondage, c'est-à-dire  $d_i = \pi_i^{-1}$ , où  $\pi_i$  est la probabilité que l'unité  $i$  soit échantillonnée. Deville et Särndal (1992) montrent que (sous la contrainte que  $G(\cdot)$  obéisse à certaines conditions), la solution du problème d'optimisation contrainte susmentionnée peut s'exprimer sous la forme :

$$w_i = d_i F(x'_i \lambda), \quad (2.2)$$

où  $F(u) = g^{-1}(u)$  désigne la fonction réciproque de  $g(u) = dG(u)/du$  et  $\lambda$  est le multiplicateur de Lagrange qui résout les équations de calage :

Deville et Särndal (1992) discutent des divers choix de la fonction  $G(\cdot)$  et de la fonction  $F(\cdot)$  connexe. Nous examinons les trois choix suivants :

*linéaire :*

$$G_L^L(u) = (1/2)(u-1)^2, \quad F_L^L(u) = 1+u;$$

*multiplicative (ratissage croisé) :*

$$G_M^M(u) = u \log(u) - u + 1, \quad F_M^M(u) = \exp(u);$$

*calage basé sur le maximum de vraisemblance :*

$$G_{ML}^{ML}(u) = u - 1 - \log(u), \quad F_{ML}^{ML}(u) = (1-u)^{-1}.$$

Voir également Deville et coll. (1993) et Fuller (2009, section 2.9) en ce qui concerne la terminologie susmentionnée pour ces fonctions. Dans le cas de la forme linéaire de  $G(\cdot)$ , le problème d'optimisation possède une solution analytique et l'estimateur par calage généralisé devient  $\bar{Y}_y = \bar{Y}_y^{yd} + (T_x - \bar{Y}_x^{yd})' \bar{B}_s$ , c'est-à-dire l'estimateur par la régression généralisée (GREG), où  $\bar{Y}_y^{yd} = \sum_s d_i y_i$ ,  $\bar{Y}_x^{yd} = \sum_s d_i x_i$  et

$$\bar{B}_s = \left( \sum_s d_i x_i x_i' \right)^{-1} \sum_s d_i x_i y_i. \quad (2.4)$$

Dans le cas de la forme multiplicative de  $G(\cdot)$ , l'estimateur calé de  $T_y$  est l'estimateur par le ratissage croisé classique (Brackstone et Rao 1979) quand  $T_x$  contient les dénombrements de population dans les catégories d'au moins deux variables auxiliaires catégoriques. Par exemple, dans le contexte de l'Enquête sur la population active du Royaume-Uni,  $x_i$  désigne le vecteur de variables indicatrices de trois variables auxiliaires catégoriques :  $x_i = (\delta_{1,1}, \dots, \delta_{A,1}, \delta_{1,2}, \dots, \delta_{B,2}, \delta_{1,3}, \dots, \delta_{C,3})'$ , où  $\delta_{a,i} = 1$  si l'unité  $i$  se trouve dans la catégorie  $a$  de la première variable auxiliaire et 0 autrement,  $\delta_{b,i} = 1$  si l'unité  $i$  est dans la catégorie  $b$  de la deuxième variable auxiliaire et 0 autrement, et ainsi de suite. Le total de population  $T_x$  de ce vecteur contient donc les dénombrements de population dans chacune des catégories (marginales) de chacune des trois variables auxiliaires. La construction des poids pour l'estimation par le ratissage croisé classique s'appuie habituellement sur l'ajustement proportionnel itératif (Brackstone et Rao 1979). Ireland et Kullback (1968) démontrent que cette méthode converge vers une solution du problème d'optimisation susmentionné.

La fonction  $G_{ML}^{ML}(u)$  mène à une version distincte basée sur le « maximum de vraisemblance » du redressement par calage, quand  $x_i$  prend la même forme désignant des variables indicatrices pour les variables auxiliaires catégoriques. Dans ce cas, la fonction objectif (2.1) peut être



# Estimation de la variance par linéarisation pour les estimateurs par calage généralisé en présence de non-réponse

Julia D'Arriago et Chris Skinner<sup>1</sup>

## Résumé

Diverses formes d'estimateurs de variance par linéarisation pour les estimateurs par calage généralisé sont définies en choisissant différents poids à appliquer a) aux résidus et b) aux coefficients de régression estimés utilisés dans le calcul des résidus. Des éléments de théorie sont présentés pour trois formes de l'estimateur par calage généralisé, à savoir l'estimateur par ratisage croisé classique, l'estimateur par calage basé sur le « maximum de vraisemblance » et l'estimateur par ratisage croisé généralisé. Une étude par simulation est effectuée en se servant des données d'une enquête sur la population active et d'une enquête sur les revenus et dépenses. Les propriétés des estimateurs sont évaluées en fonction de l'échantillonnage ainsi que de la non-réponse. L'étude révèle peu de différences entre les propriétés des divers estimateurs par calage pour un plan d'échantillonnage et un modèle de non-réponse donnés. En ce qui concerne les estimateurs de variance, l'approche consistant à pondérer les résidus par les poids de sondage peut être fortement biaisée en présence de non-réponse. L'approche de pondération des résidus par les poids calés a tendance à produire un biais nettement plus faible. Le choix de différents types de poids pour produire les coefficients de régression a peu d'incidence.

Mots clés : Calage ; non-réponse ; ratisage ; estimation de la variance ; poids.

## 1. Introduction

Dans les sondages, le recours à la pondération pour corriger le biais de non-réponse est une approche très répandue. L'estimation par calage généralisé (Deville, Särndal et Sautory 1993) fournit une classe de méthodes de pondération qui peuvent être utilisées quand les totaux de population des variables auxiliaires sont disponibles. Ces méthodes peuvent, en principe, éliminer le biais de non-réponse (en grand échantillon) quand la probabilité de non-réponse est reliée aux valeurs des variables auxiliaires par un modèle linéaire généralisé.

Dans le présent article, nous présentons certains éléments de théorie concernant l'estimation de la variance par linéarisation pour ce genre de méthodes en présence de non-réponse. Nous décrivons également une étude par simulation des propriétés de divers estimateurs par calage et des estimateurs de variance connexes dans des conditions choisies pour imiter deux enquêtes européennes réalisées par des instituts nationaux de statistiques. Nous considérons trois formes d'estimateur par calage, à savoir l'estimateur par ratisage croisé (raking ratio) classique, l'estimateur par calage du « maximum de vraisemblance » (Brackstone et Rao 1979 ; Fuller 2002) et l'estimateur par la régression généralisée (GREG). Le premier estimateur a été utilisé en pratique dans l'Enquête sur la population active (EPA) du Royaume-Uni, qui est la première enquête sur laquelle est fondée notre étude par simulation. Une version du deuxième estimateur a été utilisée en pratique dans l'Enquête sur les revenus et les dépenses (ERD) de l'Allemagne, qui est la deuxième

enquête sur laquelle s'appuie notre étude par simulation. L'estimateur GREG est d'usage très répandu dans de nombreuses enquêtes, en particulier dans le contexte de la non-réponse (Särndal et Lundström 2005). Un certain nombre de méthodes de pondération, qui n'entrent pas dans la catégorie des méthodes de calage généralisé considérées ici, ont été proposées. Voir Särndal et Lundström (2005) pour un compte rendu historique et Kott (2006), ainsi que Chang et Kott (2008) pour certains développements récents, où les variables auxiliaires pour lesquelles l'information au niveau de la population est disponible peuvent différer des variables utilisées comme covariables dans le modèle linéaire généralisé de la probabilité de non-réponse.

Le présent article porte avant tout sur l'estimation de la variance et, en particulier, sur les méthodes de linéarisation, pour lesquelles un certain nombre de formes légèrement différentes d'estimateurs de variance sont décrites dans la littérature. Dans notre étude par simulation, nous comparerons les propriétés de divers estimateurs par calage et des estimateurs de variance connexes en ce qui concerne les effets de l'échantillonnage ainsi que de la non-réponse. Une étude par simulation antérieure effectuée par Stukel, Hidiroglou et Särndal (1996) n'a révélé que peu de différences entre deux formes d'estimateur par linéarisation en ce qui concerne l'échantillonnage. Cependant, il existe des raisons pour lesquelles la non-réponse pourrait entraîner des écarts plus importants. Les conditions pour l'absence de biais dans les méthodes d'estimation par calage sous des modèles de non-réponse varient selon la méthode





- Sämdal, C.-E., Swensson, B. et Wreiman, J. (1992). *Model-Assisted Survey Sampling*. New York : Springer-Verlag.
- Sforzi, F. (1991). I distretti industriali marshalliani nell'economia italiana. Dans *Distretti industriali e cooperazione fra imprese in Italia*, (Eds., F. Pyke, G. Becattini et W. Sengenberger). Quaderni di Studi e Informazioni, 34.
- Sforzi, F., et Lorenzini, F. (2002). I distretti industriali. Dans *Ministero delle Attività Produttive-IPL, L'esperienza italiana dei distretti industriali*. Roma, IPL.
- Spiegelhalter, D.J., Best, N., Carlin, B.P. et Van der Linde, A. (2002). Bayesian measures of model complexity and fit (avec discussion). *Journal of the Royal Statistical Society, Series B*, 64, 583-639.
- Spiegelhalter, D.J., Thomas, A., Best, N.G. et Gilks, W.R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling*. Version 0.50. Medical Research Council Biostatistics Unit, Cambridge.
- Van Ophem, H. (1999). A general method to estimate correlated discrete random variables. *Econometric Theory*, 15, 228-237.
- You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, 30, 3-15.
- You, Y., Rao, J.N.K. et Gambino, J. (2003). Estimation du taux de chômage fondée sur un modèle pour l'Enquête sur la population active du Canada : une approche bayésienne hiérarchique. *Techniques d'enquête*, 29, 27-36.
- Winkelmann, R. (2003). *Econometric Analysis of Count Data*. Springer, Berlin.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York : Springer-Verlag.
- You, Y. (2008). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions intraprovinciales au Canada. *Techniques d'enquête*, 34, 1, 21-31.
- You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 107-114.

et  $\Sigma = \text{COV}(\log \tau)$ . Alors, la loi marginale de  $y$  est une loi multivariée Poisson-Log-normale (MPLN), qui est un mélange Log-normal de  $d$  lois de Poisson  $\text{Po}(\tau_j)$  indépendantes, c'est-à-dire  $y | \lambda, \Sigma \sim \text{PLN}_d(\lambda, \Sigma)$ . En désignant le  $(j, h)$ ,  $j, h = 1, 2, \dots, d$  élément de  $\Sigma$  par  $\sigma_{jh}$ , nous pouvons obtenir facilement les moments marginaux par la voie des résultats d'espérance conditionnelle et des propriétés standard des lois de Poisson et Log-normale :

$$E(y_j | \lambda, \Sigma) = \exp(\lambda_j + \sigma_{jj}/2) = \zeta_j$$
$$V(y_j | \lambda, \Sigma) = \zeta_j + \zeta_j^2 [\exp(\sigma_{jj}) - 1]$$
$$\text{COV}(y_j, y_h | \lambda, \Sigma) = \zeta_j \zeta_h [\exp(\sigma_{jh}) - 1], j \neq h.$$

Notons que le modèle MPLN tient compte de la surdispersion fournie par  $\sigma_{jj} > 0$ , ce qui mène à  $V(y_j | \lambda, \Sigma) > E(y_j | \lambda, \Sigma)$ . En outre, la structure de corrélation des dénombrements n'est pas contrainte, puisque  $\text{COV}(y_j, y_h | \lambda, \Sigma)$  peut être positive ou négative selon le signe  $\sigma_{jh}$ . Aitchison et Ho (1989), ainsi que Good et Pirog-Good (1989), ont étudié une loi MPLN bivariée, bien qu'uniquement dans les cas où il n'existerait pas de covariables. Cependant, le même modèle peut facilement être étendu en vue de prendre les covariables en considération (Chib et Winkelmann 2001).

## Bibliographie

- Aitchison, J., et Ho, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76, 643-653.
- Aroa, V., et Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.
- Baldi, C., Bellisai, D., Fivizzani, S. et Sorrentino, M. (2007). Production of job vacancy statistics: Coverage. Contribuiti Istat W. Sengenberger). Internation Labor Office, Genève.
- Becattini, G. (1992). The Marshallian industrial district as a socio-economic notion. Dans *Industrial Districts and International Co-operation in Italy*, (Eds., F. Pyke, G. Becattini et W. Sengenberger), Internation Labor Office, Genève.
- Chattopadhyay, M., Lahiri, P., Larsen, M. et Reimnitz, J. (1999). Estimation composite de la prévalence des drogues pour des zones infrats. *Techniques d'enquête*, 25, 91-97.
- Chen, S. (2001). Empirical best prediction and hierarchical Bayes methods in small area estimation. Thèse de doctorat, Department of Mathematics and Statistics, University of Nebraska, Lincoln.
- Chib, S., et Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19, 428-435.
- Cohen, M.L. (2000). Evaluation of Census Bureau's small-area poverty estimates. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 62-68.
- Statistique Canada, N° 12-001-X au catalogue
- Datta, G.S., Fay, R.E. et Ghosh, M. (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation. *Proceedings of the Census 1991 Annual Research Conference*, U. S. Bureau of the Census, Washington, DC, 63-79.
- Datta, G.S., Ghosh, M., Nangia, N. et Natarajan, K. (1996). Estimation of median income of four-person families: A Bayesian approach. Dans *Bayesian Analysis in Statistics and Econometrics*, (Eds., D.A. Berry, K.M. Chaloner et J.M. Geweke). New York : John Wiley & Sons, Inc., 129-140.
- Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 488, 1074-1082.
- Elazar, D. (2004). Small area estimation of disability in Australia. *Statistics in Transition*, 6, 5, 667-684.
- Fabrizi, E., Ferrante, M.R. et Paci, S. (2005). Estimation of poverty indicators at sub-national level using multivariate small area models. *Statistics in Transition*, 7, 3, 587-608.
- Fabrizi, E., Ferrante, M.R. et Paci, S. (2008). Measuring sub-national income poverty by using a small area multivariate approach. *Review of Income and Wealth*, 54, 4, 597-615.
- Fay, R.E. (1987). Application of multivariate regression to small domain estimation. Dans *Small Area Statistics*, (Eds., R. Platek, J.N.K. Rao, C.-E. Sæmøl et M.P. Singh). New York : John Wiley & Sons, Inc., 91-102.
- Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-511.
- Ghosh, M., Nangia, N. et Kim, D. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.
- Good, D.H., et Pirog-Good, M.A. (1989). Models for bivariate count data with an application to teenage delinquency and paternity. *Sociological Methods and Research*, 17, 4, 409-431.
- Istat (1997). I sistemi locali del lavoro 1991. *Argomenti*, Roma 1997, 10.
- Lahiri, P., et Rao, J.N.K. (1995). Robust estimation of mean square error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.
- Liu, B., Lahiri, P. et Kalton, G. (2007). Hierarchical Bayes modeling of survey weighted small area proportions. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 160-165.
- Rao, J.N.K. (2003). *Small Area Estimation*. New Jersey : John Wiley & Sons, Inc.



Tableau 3  
Vérification prédictive a posteriori : sommaires de  $p^*_j$  et de  $d^*_j$  calculés par rapport à  $i$

Modèle	Ensemble de données		$p$		$p^*_i$	$p^*_i$	$d^*_i$	$d^*_i$
MNN-MCEEL	DNN	0,034	Min.	0,000	0,003	-3,764	-2,867	
			Médiane	0,591	0,616	0,257	0,295	
			Max.	0,995	0,993	2,656	-2,515	
MNPLN-MCEEL	DNN	0,65	Min.	0,154	0,129	-0,965	-1,165	
			Médiane	0,535	0,561	0,124	0,149	
			Max.	0,891	0,912	1,216	1,286	
MNPLN-MCEEM	DNN	0,78	Min.	0,090	0,134	-1,085	-0,983	
			Médiane	0,515	0,519	-0,084	-0,085	
			Max.	0,916	0,914	1,401	1,787	
MNPLN-MCEEM	DNN+DN	0,79	Min.	0,072	0,111	-1,164	-0,945	
			Médiane	0,506	0,523	-0,076	-0,094	
			Max.	0,903	0,913	1,301	1,778	

Tableau 4  
Estimations directes et HB pour les secteurs industriels : en italique, estimations HB dont les intervalles de crédibilité couvrent les estimations directes

Estimations directes										Estimations HB									
MNN-MCEEL (DNN)					MNPLN-MCEEL (DNN)					MNPLN-MCEEM (DNN)					(DNN+DN) MNPLN-MCEEM				
Int. de crédit. à 95 %					Int. de crédit. à 95 %					Int. de crédit. à 95 %					Int. de crédit. à 95 %				
${}^A\hat{\theta}_{s1}$ se ( ${}^A\hat{\theta}_{s1}$ )	${}^A\hat{\theta}_{HB}$	${}^A\hat{\theta}_{s2}$ se ( ${}^A\hat{\theta}_{s2}$ )	${}^A\hat{\theta}_{HB}$	Int. de crédit. à 95 %	${}^A\hat{\theta}_{s1}$ se ( ${}^A\hat{\theta}_{s1}$ )	${}^A\hat{\theta}_{HB}$	${}^A\hat{\theta}_{s2}$ se ( ${}^A\hat{\theta}_{s2}$ )	${}^A\hat{\theta}_{HB}$	Int. de crédit. à 95 %	${}^A\hat{\theta}_{s1}$ se ( ${}^A\hat{\theta}_{s1}$ )	${}^A\hat{\theta}_{HB}$	${}^A\hat{\theta}_{s2}$ se ( ${}^A\hat{\theta}_{s2}$ )	${}^A\hat{\theta}_{HB}$	Int. de crédit. à 95 %	${}^A\hat{\theta}_{s1}$ se ( ${}^A\hat{\theta}_{s1}$ )	${}^A\hat{\theta}_{HB}$	${}^A\hat{\theta}_{s2}$ se ( ${}^A\hat{\theta}_{s2}$ )	${}^A\hat{\theta}_{HB}$	Int. de crédit. à 95 %
1 702,0	41,3	1 077,0	964,3	1 201,0	1 266,0	1 055,0	1 509,0	1 649,0	1 434,0	1 906,0	1 630,0	1 406,0	1 598,0	1 899,0	1 702,0	41,3	1 077,0	964,3	1 201,0
2 1 758,8	41,9	1 793,0	2 091,0	1 720,0	4 814,0	2 011,0	1 975,0	1 665,0	2 347,0	1 908,0	1 598,0	2 291,0	1 758,8	41,9	1 793,0	2 091,0	1 720,0	4 814,0	2 011,0
3 725,0	26,9	557,8	460,6	662,7	534,6	435,8	642,3	696,6	573,3	842,3	682,8	575,5	725,0	26,9	557,8	460,6	662,7	534,6	435,8
4 373,9	19,3	202,7	123,0	294,8	192,1	129,1	277,0	370,0	291,1	471,4	319,8	252,1	373,9	19,3	202,7	123,0	294,8	192,1	129,1
5 142,4	11,9	158,2	66,5	258,2	146,0	98,4	205,7	235,6	164,3	326,9	149,7	108,3	142,4	11,9	158,2	66,5	258,2	146,0	98,4
6 5 624,1	75,0	4 134,0	3 800,0	4 484,0	5 235,0	4 814,0	5 670,0	5 537,0	5 136,0	5 963,0	5 594,0	6 029,0	5 624,1	75,0	4 134,0	3 800,0	4 484,0	5 235,0	4 814,0
7 887,7	29,8	659,9	549,1	783,7	629,6	526,4	743,4	872,7	761,7	1 003,0	844,6	732,3	887,7	29,8	659,9	549,1	783,7	629,6	526,4
8 223,9	15,0	263,3	188,2	340,6	260,6	182,8	351,3	362,0	262,8	494,1	288,7	203,1	223,9	15,0	263,3	188,2	340,6	260,6	182,8
9 661,5	25,7	893,7	790,3	999,4	777,6	624,7	948,7	931,0	754,8	1 150,0	803,3	638,7	661,5	25,7	893,7	790,3	999,4	777,6	624,7
10 1 792,6	42,3	1 460,0	1 334,0	1 598,0	1 579,0	1 381,0	1 798,0	1 847,0	1 650,0	2 074,0	1 813,0	1 610,0	1 792,6	42,3	1 460,0	1 334,0	1 598,0	1 579,0	1 381,0

Remerciements

Les auteurs remercient le rédacteur en chef, le rédacteur associé et l'examinateur de leurs commentaires et suggestions utiles. Les travaux de recherche à l'origine du présent article ont été financés, en partie, par les subventions Miur-PRIN 2003/2003133249 et Miur-Prin 2008/2008CEFF37-001.

Annexe

La loi multivariée Poisson-Log-normale

Soit  $y = (y_1, y_2, \dots, y_j, \dots, y_d)$  un vecteur de dimensions  $d$ , et supposons que  $y_j | \tau_j \sim \text{Po}(\tau_j)$ , avec  $y_j | \tau_j \perp y_{j'} | \tau_{j'} (j \neq j')$ . Soit le vecteur de paramètres  $\tau = (\tau_1, \tau_2, \dots, \tau_j, \dots, \tau_d)$  qui suit une loi Log-normale multivariée, c'est-à-dire  $\tau | \lambda, \Sigma \sim \text{LN}_d(\lambda, \Sigma)$ , où  $\lambda = E(\log \tau)$

comme  $\theta_{HB}^f = T^{-1} \sum_{i=1}^T (\sum_{j=1}^J \theta_{jN}^{f,NZ})' (\theta_{jN}^{f,NZ})^*$ . Le tableau 4 donne les résultats sommatifs pour  $\theta_{jN}^f$  et  $\theta_{HB}^f$ .

Pour tous les modèles multivariés, nous avons examiné les variantes qui suivent des lois a priori : nous avons utilisé des lois a priori uniformes non informatives indépendantes pour les éléments des vecteurs  $\alpha, \beta, \gamma, \alpha, \beta^*,$  et  $\gamma^*$  :  $\sigma_{v,12}^{-2} \sim U^+, j=1, 2, \rho_v \sim U(-1,1), \sigma_{v,12} = \rho_v (\sigma_{v,12}^2 \sigma_{v,12}^2)^{1/2}$ . Nous avons fait la même chose pour les éléments de la matrice  $\Sigma^*$  dans le modèle MNN. Nous n'avons relevé aucun changement pertinent dans les lois a posteriori des paramètres d'intérêt.

**5.1 Comparaison des modèles MNPLN-MCEEL et MNN-MCEEL sur l'ensemble de données DNN**

Nous constatons que le modèle MNPLN-MCEEL sur-

passe largement le modèle MNN-MCEEL en ce qui concerne le DIC (tableau 2). Ce dernier modèle présente un manque d'ajustement, sa valeur  $p$  étant égale à 0,034 (tableau 3), tandis qu'une valeur  $p$  de 0,65 donne à penser que le modèle MNPLN-MCEEL est adéquat. Ce résultat est confirmé quand nous comparons les mesures  $p_{ij}^*$  et  $d_{ij}^*$  (tableau 3) pour les deux modèles. Dans le cas du modèle MNN-MCEEL,  $p_{ij}^*$  varie, selon le domaine, de 0,000 à 0,995 pour les nouvelles recrues NR ( $j=1$ ) et de 0,003 à 0,993 pour les recrues de substitution RS ( $j=2$ ), respectivement, ce qui témoigne d'une surestimation et d'une sous-estimation dans certains domaines. En outre, les statistiques sommatifs pour les résidus standardisés  $d_{ij}^*$  indiquent que certaines valeurs prédites se situent à plus de deux écarts-types des valeurs observées correspondantes. Les mêmes mesures pour le modèle MNPLN-MCEEL indiquent que l'ajustement est adéquat.

Nous constatons aussi que le modèle MNPLN-MCEEL surpasse le modèle MNN-MCEEL quand le rendement est évalué en fonction des estimations pour les grands domaines (tableau 4). En fait, les intervalles de crédibilité pour le modèle MNN-MCEEL couvrent seulement deux estimations directes agrégées pour les RS, tandis que sous le modèle MNPLN-MCEEL, les intervalles de crédibilité couvrent six estimations directes agrégées pour les NR et six pour les RS.

**5.2 Comparaison des modèles MNPLN-MCEEL et MNPLN-MCEEM sur l'ensemble de données DNN**

Les valeurs de  $p, p_{ij}^*$  et  $d_{ij}^*$  sont approximativement comparables pour les modèles MNPLN-MCEEL et MNPLN-MCEEM (tableau 3). De même, les estimations fondées sur un modèle produites par MNPLN-MCEEL

calculé sur ce modèle. Etant donné ces résultats, nous concluons que l'ajustement du modèle MNPLN-MCEEM est adéquat.

**5.3 Évaluation du rendement du modèle MNPLN-MCEEM sur l'ensemble de données DNN+DN**

Nous observons que le rendement du modèle MNPLN-MCEEM sur l'ensemble complet de données en ce qui concerne  $p, p_{ij}^*$  et  $d_{ij}^*$  est satisfaisante et comparable à celui du même modèle sur l'ensemble de données DNN (tableau 3). De toute évidence, les valeurs du critère DIC obtenues pour les deux modèles ne peuvent être comparées, les deux modèles étant estimés sur différents ensembles de données.

Comme le montre le tableau 4, tous les intervalles de crédibilité calculés au moyen de ce modèle couvrent des estimations directes sur de grands domaines ; autrement dit, la concordance des estimations HB avec les estimations directes est très satisfaisante. Ce résultat peut s'expliquer en observant que la probabilité de dénombrements nuls est plus grande dans les petits domaines, qui sont caractérisés par un petit nombre d'employés (la covariable dans tous les modèles). Par conséquent, le fait d'estimer les modèles sur l'ensemble de données DNN peut produire des estimations biaisées du paramètre  $\beta$ . Nous concluons que l'intégration d'un modèle de covariance d'échantillonnage dans le modèle d'estimation sur petits domaines MNPLN accroît sensiblement la fiabilité des estimations sur petits domaines. Afin de rendre compte du gain d'efficacité des estimations HB, nous avons calculé sur l'ensemble de données DNN la réduction moyenne en pourcentage du CV (You 2008), définie comme la moyenne de l'écart entre le CV direct et le CV HB (le ratio de la racine carrée de la variance a posteriori et de la moyenne a posteriori) par rapport au CV direct. La réduction moyenne du CV est de 23,1 % pour les NR et de 29,1 % pour les RS.



moienne a posteriori de la déviance ( $D$ ) et d'une mesure de complexité correspondant à la différence entre  $\bar{D}$  et la déviance évaluée à la moyenne a posteriori du paramètre. De cette façon, un modèle aura la préférence s'il donne une valeur du DIC plus faible (Spiegelhalter, Best, Carlin et Van der Linde 2002).

Afin de vérifier la force de l'approche multivariée de l'estimation sur petits domaines, nous utilisons comme référence les versions univariées des modèles discutés aux sections 4.2 et 4.3, définies comme il suit. Pour tous les modèles, nous posons que  $\sigma_{v,12} = 0$  dans  $\Sigma_v$ , et nous supposons que  $\sigma_{v,11} \perp \sigma_{v,22}$ ,  $\sigma_{v,11}^{(j)} \sim U(0, U^{(j)})$ ,  $j = 1, 2$ . Pour les modèles MCEEL, nous posons que  $\Psi_i = \text{diag}(\Psi_i)$ , tandis que pour les modèles MCEEM, nous posons que  $\sigma_{v,12} = 0$  dans (5). En outre, nous obtenons un nouvel ensemble d'estimations pour les paramètres  $\bar{\theta}_{11}$  et  $\bar{\theta}_{22}$  en posant que  $\bar{p}_k = 0$  dans le modèle de la section 4.1.

Le tableau 2 donne les valeurs du critère DIC pour l'ensemble complet de modèles d'estimations sur petits

**Tableau 2**  
Comparaison des modèles en utilisant la statistique DIC

Modèle	Ensemble de données	DIC
MNN-MCEEL (version univariée)	DNN	2 742,2
MNPLN-MCEEL (version univariée)	DNN	2 745,4
MNPLN-MCEEL (version univariée)	DNN	2 656,9
MNPLN-MCEEL (version univariée)	DNN	2 661,0
MNPLN-MCEEM (version univariée)	DNN	2 623,6
MNPLN-MCEEM (version univariée)	DNN	2 638,1
MNPLN-MCEEM (version univariée)	DNN+DN	3 202,7
MNPLN-MCEEM (version univariée)	DNN+DN	3 214,3

Tous les modèles multivariés considérés donnent de meilleurs résultats en ce qui concerne le DIC que leurs analogues univariés (tableau 2). En outre, pour tous les modèles multivariés, nous constatons que les intervalles de crédibilité a posteriori de  $p_v = \sigma_{v,12} / \sqrt{\sigma_{v,11}\sigma_{v,22}}$  ne contiennent pas la valeur zéro. Par conséquent, dans les paragraphes qui suivent, nous nous concentrons sur les modèles multivariés. Nous avons vérifié l'adéquation des modèles multivariés spécifiés par des vérifications prédictives a posteriori. Nous avons généré les valeurs simulées d'une mesure de divergence appropriée à partir de la loi prédictive a posteriori et nous les avons comparées aux valeurs de la même mesure calculée d'après les données observées. Soit  $\theta_{\text{obs}}$  et  $\theta_{\text{sim}}$  les données observées et générées, respectivement. La valeur  $p$  prédictive a posteriori est définie comme  $p = P\{d(\theta_{\text{sim}}, \theta) > d(\theta_{\text{obs}}, \theta) | \theta_{\text{obs}}\}$ . Nous envisageons une mesure de divergence proposée dans Datta et coll. (1999), définie comme suit :

$$d(\theta, \theta) = \sum_{i=1}^I (\theta_i - \theta_i)' \Psi^{-1} (\theta_i - \theta_i). \quad (6)$$

Le calcul de la valeur  $p$  au moyen du résultat de la simulation MCMC est simple. Les valeurs extrêmes de la probabilité  $p$  indiquent un manque d'ajustement d'un modèle donné. À l'exemple de Rao (2003, pages 245-246) et de You et Rao (2002), nous calculons deux statistiques utiles pour évaluer l'ajustement d'un modèle au niveau du domaine individuel. La première statistique  $p_{ij}^* = P(\theta_{ij, \text{sim}} < \theta_{ij, \text{obs}} | \theta_{\text{obs}})$ , renseigne sur le degré de surestimation ou de sous-estimation systématique de  $\theta_{ij, \text{obs}}$ . La deuxième statistique est définie ainsi :

$$d_{ij}^* = [E(\theta_{ij} | \theta_{\text{obs}}) - \theta_{ij, \text{obs}}] / \sqrt{V(\theta_{ij} | \theta_{\text{obs}})},$$

où l'espérance et la variance sont calculées en fonction de la loi prédictive a posteriori. Le tableau 3 résume les résultats relatifs à  $p$ ,  $p_{ij}^*$  et  $d_{ij}^*$ .

En guise de vérification supplémentaire de la cohérence des données, nous avons calculé les estimations directes et celles fondées sur le modèle de  $\theta_{ij, s}$ ,  $s = 1, \dots, 10$ , c'est-à-dire le nombre total de NR et de RS pour les dix domaines déterminés en classant les entreprises uniquement selon le secteur industriel. Soit  $w_{is} = 1$  si le nombre de recrues dans le domaine  $i$  se rapporte au secteur industriel  $s$  et  $w_{is} = 0$  ; autrement, alors :

$$\theta_{ij, s} = \sum_{i=1}^I \theta_{ij} w_{is}. \quad (7)$$

À ce niveau d'agrégation, les estimations directes peuvent être considérées comme exactes. Par conséquent, étant donné deux ensembles d'estimations fondées sur un

modèle agrégé sont calculées en se basant sur les données de sortie MCMC. Pour les modèles ayant trait aux données DNN, nous avons effectué l'agrégation selon (7) à chaque étape  $t$ ,  $t = 1, \dots, T$ , de la simulation MCMC, avec les échantillons  $\theta_{ij}^*$  et  $\theta_{ij}^{**}$  tirés respectivement de la loi a posteriori de  $\theta_{ij}$  pour les domaines appartenant à l'ensemble DNN et de la loi prédictive de  $\theta_{ij}$  pour les domaines appartenant à l'ensemble DN. L'estimateur HB est défini comme  $\hat{\theta}_{\text{HB}} = T^{-1} \sum_{t=1}^T (\sum_{i \in \text{VZC}} \theta_{ij}^* w_{it} + \sum_{i \in \text{ZC}} \theta_{ij}^{**} w_{it})$ . Sinon, pour le modèle sur les données DNN+DN, nous avons agrégé selon (7) les échantillons MCMC des lois a posteriori de  $\theta_{ij}$ . Dans ce cas, l'estimateur HB est défini



fait, (3) ne garantit pas la positivité de  $\theta_i$  ni, par conséquent, des éléments diagonaux de  $\Psi_i$ .

## 5. Analyse des données

À la section 5.1, nous comparons le modèle MNPLN au modèle MNN de référence et à leurs analogues univariés. Pour les deux modèles, nous supposons que les matrices de covariance des erreurs d'échantillonnage sont lissées (MCEEL); nous désignons donc les deux stratégies par MNPLN-MCEEL et MNN-MCEEL dans la suite. Puisque ces modèles ne nous permettent pas de résoudre le problème des dénombrements nuls, nous faisons référence dans cette analyse à l'ensemble de données DNN. À la section 5.2, nous comparons la stratégie intégrée d'estimation sur petits domaines basée sur le modèle MNPLN et la matrice de covariance des erreurs d'échantillonnage fondée sur un modèle MCEEM (MNPLN-MCEEM), que nous avons présentées à la section 4.3, avec la stratégie basée sur le MNPLN-MCEEL. Nous limitons l'analyse à l'ensemble de données DNN afin d'évaluer les deux stratégies sous les mêmes conditions. Enfin, à la section 5.3, nous évaluons le rendement global du modèle d'estimation sur petits domaines proposé MNPLN-MCEEM pour l'ensemble complet de données (DNN+DN).

Pour tous les modèles, nous avons obtenu les lois a posteriori des paramètres par intégration Monte Carlo au moyen de l'algorithme d'échantillonnage de Gibbs. Nous avons utilisé le logiciel de simulation MCMC WinBUGS (Spiegelhalter, Thomas, Best et Gilks 1995) pour exécuter trois chaînes parallèles (chacune comportant 25 000 exécutions) en tirant le point de départ d'une distribution surdispersée. Les codes de WinBUGS sont disponibles à l'adresse URL <http://www2.stat.unibo.it/trivisano/>. Nous avons surveillé la convergence de l'échantillonneur de Gibbs par inspection visuelle des graphiques des chaînes et des diagrammes d'autocorrélation, ainsi qu'au moyen du facteur de réduction d'échelle potentielle proposé par Gelman et Rubin (1992). Bien que la convergence ait été rapide pour tous les modèles, nous avons écarté les 5 000 premières itérations de chaque chaîne. Dans les modèles multivariés, l'autocorrélation relativement forte des chaînes est réduite en amincissant la chaîne (une valeur sur trois a été considérée pour les sommaires a posteriori). Voir Rao (2003, pages 228-232) pour des précisions.

Les propriétés des modèles d'estimation sur petits domaines dont nous avons discuté aux sections 4.2 et 4.3 sont comparées en utilisant diverses mesures. Afin de choisir parmi les modèles concurrents, nous avons calculé le critère d'information de déviance (DIC pour *Deviance Information Criterion*). Le DIC est un critère de sélection de modèle en vertu duquel le rendement d'un modèle est évaluée comme la somme d'une mesure d'ajustement (la

d'information a priori au sujet des paramètres du modèle, donc pour établir une spécification diffuse, mais correcte, des priors. Les moyennes a posteriori  $\theta_{HB}^i = E(\theta_i | \Psi_i)$  sont prises comme estimateurs des paramètres de petit domaine, tandis que la variance a posteriori  $V(\theta_i | \Psi_i)$  est utilisée comme mesure de l'incertitude.

Par souci de comparaison, nous prenons comme référence le modèle multivarié Normal-Normal (MNN) standard, dans lequel le modèle d'échantillonnage est défini comme en (1) et le modèle de lien, comme suit :

$$(3) \quad \theta_i \sim \text{ind } N_2(\mu_i^*, \Sigma_i^*),$$

où  $\mu_i^* = \alpha^* + \gamma^* Z_i + \beta^* Z_i x_i^*$ . Les paramètres  $\alpha^*, \gamma^*, \beta^*$  et leurs lois a priori sont définis comme  $\alpha, \gamma$  et  $\beta$  dans le modèle précédent.

## 4.3 Modèle d'estimation sur petits domaines MNPLN

À fin de tenir compte de la variabilité supplémentaire due aux matrices de covariance estimées des erreurs d'échantillonnage, ainsi que pour surmonter le problème des variances et covariances nulles, nous proposons une solution dans l'esprit de celle proposée par Arora et Lahiri (1997), Liu et coll. (2007) ainsi que You (2008). Nous intégrons le modèle pour les matrices de covariance des erreurs d'échantillonnage de la section 4.1 dans les modèles d'estimation sur petits domaines (1) et (2). Donc, ici, nous faisons référence à l'ensemble complet de 208 domaines.

Dans ce contexte, le modèle d'échantillonnage sur petits domaines est formulé comme d'habitude, c'est-à-dire  $\theta_i | \Psi_i \sim \text{ind } N_2(\theta_i^*, \Psi_i^*)$ ,  $i = 1, \dots, 208$ . Conformément aux hypothèses  $Y_i^*$  relatives aux équations formulées à la section 4.1, en supposant que les  $Z_i^*$  sont connues et en supposant que  $\theta_{ij}^* = N_{ij}^*$ , les éléments de la matrice de covariance des erreurs d'échantillonnage  $\Psi_i^*$  peuvent être exprimés de la façon suivante :

$$(4) \quad \Psi_{i,jj}^* = K[\theta_{ij}^*/N_i^* + \theta_{ij}^{*2}/N_i^{*2}(\exp(\hat{\theta}_{ij}^{*2}/N_i^{*2}) - 1)]$$

$$(5) \quad \Psi_{i,12}^* = K[N_i^{*2}\theta_{i1}^*\theta_{i2}^*(\exp(\hat{\theta}_{i12}^{*2}/N_i^{*2}) - 1)]$$

où  $\hat{\theta}_{ij}^*$ ,  $j = 1, 2$  et  $\hat{\theta}_{i12}^*$  sont les moyennes a posteriori des paramètres  $\theta_{ij}^*$  et  $\theta_{i12}^*$ , respectivement, calculées, en utilisant le modèle de la section 4.1.

Puisque les matrices de covariance des erreurs d'échantillonnage sont exprimées en fonction des paramètres  $\theta_i$ , elles peuvent être considérées ici comme des matrices de covariance des erreurs d'échantillonnage fondées sur un modèle (MCEEM). Les moyennes a posteriori  $\theta_{HB}^i = E(\theta_i | \Psi_i)$  sont utilisées comme estimateurs des  $\theta_i^*$ , tandis que la variance a posteriori  $V(\theta_i | \Psi_i)$  est utilisée comme mesure de l'incertitude.

Nous notons que le modèle MNN ne peut pas être mis en œuvre en suivant l'approche intégrée décrite plus haut. En

où  $\mathbf{I}_i = E(C_i | n_i, \mathbf{I}_i)$ ,  $i = 1, 2, \dots, 158$ , et les éléments  $(j, h)$  de  $C_i$  sont définis comme  $C_{i,jh} = n_i^{-1} \sum_{t=1}^{n_i} (Y_{ijt} - \bar{Y}_{ijh})^2$ , où  $\bar{Y}_{ijh} = n_i^{-1} \sum_{t=1}^{n_i} Y_{ijt}$ .

Si les paramètres  $\zeta_{ij}$  sont connus, alors  $E(C_i | n_i, \mathbf{I}_i)$  dépend uniquement des éléments de la matrice  $\Sigma_i$ . Nous proposons d'estimer  $\zeta_{ij}$  en utilisant l'estimateur fondé sur le plan  $\hat{\zeta}_{ij} = N_i^{-1} \theta_{ij}^T$ . Donc, nous pouvons exprimer chaque élément de la matrice  $\mathbf{I}_i$  comme une fonction des estimations  $\hat{\zeta}_{ij}$  et des éléments de la matrice  $\Sigma_i$  :

$$\begin{aligned} \Gamma_{i11} &= \hat{\zeta}_{i1}^2 + \hat{\zeta}_{i2}^2 (\exp(\sigma_{i11}^2) - 1) \\ \Gamma_{i22} &= \hat{\zeta}_{i2}^2 + \hat{\zeta}_{i2}^2 (\exp(\sigma_{i22}^2) - 1) \\ \Gamma_{i12} &= \hat{\zeta}_{i1} \hat{\zeta}_{i2} (\exp(\sigma_{i12}^2) - 1) \end{aligned}$$

où  $\sigma_{i11} = \sigma_{i1}^T \mathbf{Z}_i \sigma_{i1} = \sigma_{i22}^T \mathbf{Z}_i \sigma_{i22} = \sigma_{i12}^T \mathbf{Z}_i \sigma_{i12}$ , étant donné que  $\mathbf{Z}_i$  est un vecteur  $3 \times 1$  de variables indicatrices indiquant la catégorie de taille de l'entreprise dans le domaine  $i$ , quant la catégorie de taille de l'entreprise dans le domaine  $i$ , et :

$$\bar{\sigma}_{i1} = \begin{pmatrix} \bar{\sigma}_{i11} \\ \bar{\sigma}_{i21} \\ \bar{\sigma}_{i31} \end{pmatrix}, \bar{\sigma}_{i2} = \begin{pmatrix} \bar{\sigma}_{i12} \\ \bar{\sigma}_{i22} \\ \bar{\sigma}_{i32} \end{pmatrix}, \bar{\sigma}_{i12} = \begin{pmatrix} \bar{\sigma}_{i112} \\ \bar{\sigma}_{i212} \\ \bar{\sigma}_{i312} \end{pmatrix}$$

c'est-à-dire que nous supposons que les paramètres  $\Sigma_i$  sont égaux pour les domaines appartenant à la même catégorie de taille d'entreprise.

Nous estimons les paramètres  $\bar{\sigma}_{i1}$ ,  $\bar{\sigma}_{i2}$ ,  $\bar{\sigma}_{i12}$  sur les données de l'ensemble DNN. Puisque nous suivons une approche bayésienne, nous devons spécifier des priors pour  $\bar{\sigma}_{k,j}$  et  $\bar{\sigma}_{k,12}$ ,  $k = 1, 2, 3$ . Nous utilisons les spécifications de priors suivantes :  $\bar{\sigma}_{k,11}^{1/2} \sim U^+$ ,  $\bar{\sigma}_{k,22}^{1/2} \sim U^+$ ,  $\bar{p}_k \sim U(-1, 1)$ , où  $\bar{\sigma}_{k,12} = \bar{p}_k (\bar{\sigma}_{k,11} \bar{\sigma}_{k,22})^{1/2}$  et  $U^+$  désignent une loi uniforme sur un sous-ensemble de  $R^+$  dont la longueur est grande mais finie. À la section 4.3, nous montrons comment ces estimations peuvent être utilisées pour intégrer le modèle d'estimation sur petits domaines à un modèle pour les matrices de covariance des erreurs d'échantillonnage.

#### 4.2 Un modèle Normal-Poisson-Log-normal

##### multivarié d'estimation sur petits domaines

À la présente section, nous proposons un modèle multivarié d'estimation sur petits domaines basé sur la loi MPLN afin d'estimer conjointement les nombres de RS et de NR en utilisant l'ensemble DNN.

Soit  $\theta_i = (\theta_{i1}, \theta_{i2})^T$  le vecteur des deux paramètres d'intérêt pour le  $i^{\text{e}}$  domaine dans l'ensemble de données DNN ( $i = 1, \dots, 158$ ), et soit  $\theta_i$  le vecteur correspondant d'estimations directes. Le modèle d'estimation sur petits domaines est constitué de deux modèles distincts. Le premier est un modèle d'échantillonnage :

$$\theta_i | \theta_i \sim \text{ind } N_2(\theta_i | \Psi_i), \quad i = 1, \dots, 158. \quad (1)$$

Comme dans Lahiri et Rao (1995), nous justifions l'hypothèse de normalité dans (1) en utilisant l'argument de la limite centrale. Il est d'usage, en pratique, de supposer

La deuxième composante du modèle d'estimation sur petits domaines est un modèle de lien qui relie  $\theta_i$  aux données auxiliaires propres au domaine :

$$\theta_i \sim \text{ind PLN}_2(\eta_i, \Sigma_i), \quad i = 1, \dots, 158,$$

où  $\eta_i = \alpha + \gamma \mathbf{Z}_i + \beta \mathbf{Z}_i x_i$ .  $\mathbf{Z}_i$  est un vecteur  $3 \times 1$  de variables indicatrices indiquant la catégorie de taille de l'entreprise dans le domaine  $i$  et  $x_i = \log(x_i^*)$ , où  $x_i^*$  est le nombre d'employés dans le domaine  $i$ .

En fin ce compte,  $\Sigma_i$  est la matrice de covariance reliée aux effets aléatoires propres au domaine :

$$\Sigma_i = \begin{pmatrix} \sigma_{i,11} & \sigma_{i,12} & \sigma_{i,22} \\ \sigma_{i,12} & \sigma_{i,21} & \sigma_{i,22} \end{pmatrix}$$

et

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ 0 & \gamma_{22} & \gamma_{23} \end{pmatrix}, \beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix}.$$

À partir d'ici, nous donnerons à ce modèle d'estimation sur petits domaines le nom de modèle « multivarié Normal-Poisson-Log-normal » (MNPLN).

Nous adoptons une approche entièrement hiérarchique bayésienne. Dans ce cadre, des modèles relativement complexes (par exemple multivariés) peuvent être facilement appliqués ; en outre, on peut approximer les lois a posteriori en utilisant des algorithmes MCMC. Le calcul d'estimations multivariées sur petits domaines, et des estimations de leur EQM en particulier, peut être difficile dans une approche fréquentiste. La spécification des priors pour le modèle décrit est la suivante :

$$\begin{aligned} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} &\sim N_2(\mathbf{0}, \mathbf{aI}_2), \\ \begin{pmatrix} \gamma_{1k} \\ \gamma_{2k} \end{pmatrix} &\sim N_2(\mathbf{0}, g_k \mathbf{I}_2) \quad k' = 2, 3, \\ \begin{pmatrix} \beta_{1k} \\ \beta_{2k} \end{pmatrix} &\sim N_2(\mathbf{0}, b_k \mathbf{I}_2) \quad k = 1, 2, 3, \\ \Sigma_i^{-1} &\sim W(s, \mathbf{I}_2), \end{aligned}$$

où  $s = 3$  et  $a, g_k, b_k$  sont grands comparativement à l'échelle des données. Il est ainsi pour refléter le manque



Tableau 1  
Variables définissant les domaines d'intérêt

LLMA regroupées selon la spécialisation de production			Taille de l'entreprise <sup>(b)</sup>	Secteur industriel <sup>(a)</sup>
<i>District Industrie<sup>(c)</sup></i>			1 à 9	1 Aliments, boissons et tabac
Aliments, boissons et tabac			10 à 49	2 Textiles et habillement
Textiles et habillement			50 à 99	3 Produits du papier, imprimerie et édition
Produits du papier, imprimerie et édition			≥ 100	4 Machinerie
Machinerie				5 Produits chimiques et métaux de base
Bijoux, instruments de musique, jeux, etc.				6 Cuir et chaussures
Cuir et chaussures				7 Bois, mobilier et équipement ménager
Bois, mobilier et équipement ménager				8 Bijoux, instruments de musique, jeux, etc.
<i>LLMA non définies comme un district<sup>(c)</sup></i>				9 Constructeurs, entrepreneurs
Fabrication non spécialisée				10 Autres activités de fabrication
Non spécialisée, à l'exclusion de la fabrication				
Tourisme				
Villes				

- (a) Tel que défini par la classification au niveau à deux chiffres de l'ATECO 91-CITI 3 et par Sforzi (1991).  
(b) Définie en fonction du nombre d'employés.  
(c) Défini conformément à Isat (1997).

#### 4. Un modèle multivarié intégré d'estimation sur petits domaines pour les données de comptage

Les données de comptage multivariées peuvent avoir une structure de corrélation non négligeable. En général, la modélisation de cette structure a une incidence importante sur l'efficacité des estimateurs et sur le calcul des erreurs-types correctes. Un certain nombre de modèles multivariés pour données de comptage ont été proposés dans la littérature, comme le modèle de Poisson multivarié, le modèle binominal négatif multivarié et les modèles de mélange Poisson et Gamma multivariés (pour une revue de ces modèles, voir Winkelmann 2003). Malheureusement, ces lois ne conviennent pas pour modéliser nos données, puisqu'elles sont basées sur l'hypothèse que la corrélation est due à un facteur individuel qui ne varie pas d'un résultat à l'autre, ce qui implique une structure de covariance restreinte à des corrélations non négatives. Dans le cas bivarié, une structure de covariance plus souple est fournie par le modèle de mélange Poisson et Normale latente (van Ophem 1999) ; cependant, toute extension à des données multivariées de plus grande dimensionnalité semble difficilement applicable.

Aitchison et Ho (1989) ont proposé une loi *d*-variée qui permet une structure de covariance non contrainte, à savoir la loi multivariée Poisson-Log-normale (MPLN). Aucune forme analytique n'existe pour cette dernière, mais elle peut être représentée par un mélange plus simple qui permet l'estimation des paramètres selon la méthode de Monte Carlo appliquée aux chaînes de Markov (MCMC) (Chib et Winkelmann 2001). Des renseignements concernant la loi MPLN sont fournis en annexe.

#### 4.1 Lissage des matrices de covariance d'échantillonnage

Comme nous l'avons mentionné plus haut, l'approche des fonctions de variance généralisées (FVG) est généralement adoptée pour traiter l'instabilité des erreurs-types dans l'estimation sur petits domaines. À la présente section, nous présentons un modèle FVG avec une fonction de régression inspirée de la loi MPLN.

Soit  $\mathbf{y}_i = [y_{i1}, y_{i2}]$  le vecteur des deux variables de résultat se rapportant à la  $i^{\text{e}}$  unité dans le  $i^{\text{e}}$  domaine. Soit  $\mathbf{y}_i | \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i \perp \mathbf{y}_i | \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i$  et  $\mathbf{y}_i | \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i \sim \text{PLN}_2(\boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i)$ . Sur la base de ces hypothèses, les moments aboutissant au deuxième ordre peuvent être exprimés de la façon suivante :

$$E(y_{ij} | \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i) = \exp(\lambda_{ij}) + \sigma_{i,jj} / 2 = \zeta_{ij} \\ V(y_{ij} | \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i) = \zeta_{ij}^2 + \sigma_{i,jj}^2 [\exp(\sigma_{i,jj}^2) - 1] \\ \text{COV}(y_{ij}, y_{ih} | \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i) = \zeta_{ij} \zeta_{ih} [\exp(\sigma_{i,jh}^2) - 1], j \neq h$$

où  $\sigma_{i,jh}$  désigne l'élément  $(j, h)$ ,  $j, h = 1, 2$  de  $\boldsymbol{\Sigma}_i$ . Pour résoudre le problème du lissage des matrices de covariance, Otto et Bell (1995) ont proposé une approche fondée sur l'hypothèse d'une loi de Wishart ; plus précisément, ils ont utilisé des estimations lissées dans un modèle normal-normal d'estimation sur petits domaines. Dans le même esprit, nous proposons une approche bayésienne s'appuyant sur la stratégie FVG qui suit. Sous échantillonnage aléatoire simple, supposons que la matrice de covariance d'échantillonnage dans le domaine  $i$ ,  $\mathbf{C}_i$  suit une loi de Wishart avec  $n_i - 1$  degrés de liberté :

$$\mathbf{C}_i | n_i, \boldsymbol{\Gamma}_i \sim W_2(n_i - 1, \boldsymbol{\Gamma}_i)$$



du processus de production locale, une meilleure subdivision territoriale serait celle des zones locales de marché du travail (*Local Labour Market Areas* ou LLMA), conformément à la définition de l'OCDE. Les LLMA sont des groupes de municipalités dans lesquelles les conditions du marché du travail sont les mêmes (pour l'emplacement des LLMA en Italie, voir Sforzi 1991). En Italie, selon la stratégie proposée par Sforzi et Lorenzini (2002) et adoptée par l'Institut italien de statistique (ISTAT), certaines LLMA sont appelées « districts industriels » (DI). Les DI sont des systèmes de production définis géographiquement caractérisés par une spécialisation dominante. Au cours des années 1990, ces districts étaient considérés comme le principal aiguillon de la croissance du système économique italien (Becattini 1992).

L'estimation du nombre de recrues de substitution et de nouvelles recrues dans les entreprises exploitées à l'intérieur et à l'extérieur des DI peut nous aider à vérifier si les districts industriels sont encore une source de dynamisme de l'économie italienne dans son ensemble. Afin de faire référence aux types de districts industriels, nous les regroupons en fonction de leur spécialisation de production. De même, Nous classons les LLMA non considérées comme des DI en fonction de leur vocation économique (les LLMA peuvent être caractérisées par une activité manufacturière particulière, une région touristique, une ville, etc.). En outre, sur le plan économique, la comparaison entre les entreprises situées dans les DI et celles non situées dans les DI est logique si le secteur industriel et la taille des entreprises sont également pris en compte. Enfin, comme nous l'avons déjà mentionné, les domaines d'intérêt sont définis par recoupement : i) des groupes de LLMA obtenus en fonction de leur spécialisation de production, ii) du secteur industriel des entreprises et iii) de la taille de l'entreprise.

Le présent article porte sur le secteur de la fabrication qui caractérise l'activité économique des districts industriels. L'analyse est limitée à deux régions de l'Italie contenant une grande quantité de districts industriels, à savoir la Toscane et l'Emilie-Romagne, et aux entreprises ayant moins de 100 employés (puisque des recensements sont effectués pour les autres catégories de taille). La population cible comprend 54 089 entreprises employant, en tout, 809 059 personnes.

### 3. Estimations directes

de l'exposé, nous évitons d'utiliser dans la mesure du même strate planifiée. Par souci de simplicité, dans la suite possible l'indice inférieur de strate.

Soit  $\theta_{ji}$  et  $\theta_{j2}$  les nombres réels de NR et de RS pour le domaine  $i$  ( $i = 1, \dots, 208$ ), respectivement. Nous définissons d'abord un estimateur direct de  $\theta_{ji}(i = 1, \dots, 208; j = 1, 2)$ . Soit  $y_{ji}^{(i)}$  la réponse de la  $i^{\text{e}}$  unité relative à la  $j^{\text{e}}$  variable dans le  $i^{\text{e}}$  domaine ( $i = 1, \dots, n_i$ , où  $n_i$  est la taille d'échantillon dans le domaine  $i$ ;  $i = 1, \dots, 208; j = 1, 2$ ). En tant qu'estimateur (direct) fondé sur le plan de sondage, nous utilisons un estimateur par le ratio au niveau du domaine défini comme étant  $\hat{\theta}_{ji}^{(i)} = \sum_{i=1}^{n_i} y_{ji}^{(i)} / (n_i / N_i)$  où  $N_i$  et  $n_i$  sont, respectivement, la taille de population et la taille de l'échantillon se rapportant au domaine  $i$ , et  $N_i' = n_i / n_{i2}$ , où  $N_{i2}$  et  $n_{i2}$  sont, respectivement, la taille de population et la taille d'échantillon de la strate  $i$  contenant le domaine  $i$  (Sämdal, Swensson et Wretman 1992, page 391).

Puisque nous estimons le nombre d'occurrences d'événements rares, dans 50 des 208 domaines, les estimations directes des nombres de NR et/ou de RS sont nulles, c'est-à-dire  $\theta_{ji} = 0$  et/ou  $\theta_{j2} = 0$ . Des estimations ponctuelles nulles impliquent que  $P(\theta_{ji}) = 0$  et/ou  $P(\theta_{j2}) = 0$ , où  $P(\theta_{ji})$  et  $P(\theta_{j2})$  sont les estimations habituelles de variance par rapport au plan de sondage de  $\theta_{ji}$  et  $\theta_{j2}$ , respectivement. Ce résultat donne une fausse impression de grande précision, alors que, dans le contexte des petits domaines, c'est plus probablement l'exact opposé qui est vrai. En outre, des estimations fondées sur le plan de sondage des nombres de NR et/ou de RS qui sont égales à zéro produisent  $C\hat{O}V(\hat{\theta}_{ji}, \hat{\theta}_{j2}) = 0$ , où  $C\hat{O}V(\hat{\theta}_{ji}, \hat{\theta}_{j2}) = 0$  est l'estimation standard fondée sur le plan de sondage de la covariance sous le plan entre  $\hat{\theta}_{ji}$  et  $\hat{\theta}_{j2}$ . Par conséquent, les covariances doivent également être lissées dans un modèle multivarié d'estimation sur petits domaines.

Dans la suite de l'exposé, l'ensemble des 50 petits domaines dont la variance estimée, ou la covariance, ou les deux sont nulles est nommé « ensemble à dénombrément nul » (DN). L'ensemble complémentaire de 158 domaines pour lesquels  $P(\theta_{ji}) > 0$  et  $P(\theta_{j2}) > 0$ , est nommé « ensemble à dénombrément non nul » (DNN). Compte tenu du processus de génération des données et de la nature des variables de résultat, nous nous attendons à observer principalement des corrélations négatives entre  $\theta_{ji}$  et  $\theta_{j2}$ . En bref, nous avons besoin, pour le lissage des matrices de covariance ainsi que pour la modélisation des paramètres de petits domaines, d'une loi appropriée qui permet l'existence d'une matrice de covariance non contrainte, c'est-à-dire des corrélations positives ainsi que négatives.

petits domaines linéaire, en écartant simplement les estimations nulles. Bien que cette solution permette de contourner le problème des « variances nulles », elle produit aussi des estimations biaisées et néglige une partie de l'échantillon.

Afin de traiter l'instabilité des estimateurs des variances et covariances, ainsi que le problème des variances d'échantillonnage estimées nulles, nous proposons une approche « intégrée » dans l'esprit de celle proposée par Arora et Lahiri (1997), Liu et coll. (2007) et You (2008). Dans un cadre hiérarchique bayésien, nous modélisons conjointement les paramètres d'intérêt et les matrices de covariance des erreurs d'échantillonnage en adoptant une solution de lissage des covariances basée une fois de plus sur le modèle de mélange Poisson-Log-normale.

Le plan de l'article est le suivant. À la section 2, nous décrivons l'ensemble de données utilisé, et à la section 3, nous présentons l'estimation directe de domaine et ses variances et covariances d'erreurs d'échantillonnage. À la section 4, nous décrivons le modèle d'estimation sur petits domaines multivarié que nous proposons pour estimer les dénombrements, ainsi que la solution recommandée pour surmonter le problème d'instabilité des estimateurs des variances et covariances d'erreurs d'échantillonnage en présence de dénombrements nuls. À la section 5, nous présentons les résultats obtenus en mesurant les propriétés du modèle retenu d'estimation sur petits domaines. Des détails sur la loi multivariée Poisson-Log-normale sont fournis en annexe.

## 2. L'Enquête Excelsior

L'Enquête Excelsior est, en Italie, l'une des sources statistiques les plus complètes de données sur la demande de main-d'œuvre, fournissant des estimations du nombre de personnes recrutées par les entreprises italiennes. Chaque année, un échantillon aléatoire simple stratifié d'environ 100 000 entreprises ayant au moins un employé sont contactées afin de leur demander le nombre de personnes qu'elles prévoient embaucher dans le court terme. Les facteurs utilisés pour la stratification sont le secteur industriel et la catégorie de taille de l'entreprise. La répartition de l'échantillon entre les strates satisfait à une contrainte imposée à la valeur maximale de l'erreur-type estimée pour un seuil de signification de 95 % (Baldi, Bellisai, Fivizzani et Sorrentino, 2007). En étant axée sur le niveau de détail géographique local, l'enquête est conçue pour produire des estimations fiables pour les provinces administratives (NUTS3, conformément à la Nomenclature des unités territoriales statistiques, à l'adresse <http://europa.eu.int/comm/eurostat/ramon/nuts>). Cette unité géographique, choisie sur la base de critères administratifs, ne semble pas être le choix idéal pour analyser la dynamique de la demande locale de main-d'œuvre. Afin d'apporter des éclaircissements sur les signes de réorganisation

le modèle univarié de Fay-Herriot (Fay et Herriot 1979), tous les articles susmentionnés reposent sur l'hypothèse de l'utilisation de l'échantillonnage normal dans les petits domaines et de modèles de lien.

Puisque les corrélations d'échantillonnage entre les estimateurs des nombres de RS et de NR sont principalement négatives, nous proposons un modèle d'estimation sur petits domaines basé sur la loi multivariée Poisson-Log-normale (MPLN). Contrairement à d'autres lois multivariées pour les données de comptage proposées dans la littérature, cette loi particulière permet des corrélations non contraintes, c'est-à-dire positives ainsi que négatives (Aitchison et Ho 1989).

Nous traitons également de la stabilité des estimateurs des variances et covariances des erreurs d'échantillonnage. Une estimation approximativement sans biais de la variance des estimateurs directs est habituellement disponible dans le contexte de l'estimation sur petits domaines. Cependant, dans le cas des modèles au niveau du domaine, il est habituel de supposer que la variance d'échantillonnage est connue et est égale à son estimation (Rao 2003, page 76). Cette hypothèse est habituellement énoncée et largement acceptée dans le cas des grands échantillons, tandis que l'estimateur de variance ainsi que les estimateurs ponctuels directs souffrent d'instabilité dans le cas des petits échantillons. En guise de solution partielle, les estimations de la variance d'échantillonnage sont souvent lissées suivant l'approche des fonctions de variance généralisées (FVG) (Wolter 1985). Dans You, Rao et Gambino (2003), les variances et covariances d'échantillonnage ont été lissées sur les domaines et au cours du temps. Afin de tenir compte de la variabilité supplémentaire associée aux variances d'échantillonnage estimées, Arora et Lahiri (1997) ont proposé une approche de lissage hiérarchique bayésienne (HB) intégrée pour les données continues. Voir You et Chapman (2006), Liu, Lahiri et Kalton (2007) et You (2008) pour diverses extensions des travaux d'Arora et Lahiri (1997).

La rareté des personnes recrutées dans certains domaines pose un problème supplémentaire qui est lié à l'instabilité des estimateurs des variances et covariances des erreurs d'échantillonnage. Quand les estimations directes de RS ou de NR (ou des deux) sont nulles, les variances et covariances estimées des erreurs d'échantillonnage sont également nulles. Notons que l'observation de variances estimées nulles ne signifie pas nécessairement que les estimations ont un haut degré d'exactitude. Cette question a été soulevée dans le cas de problèmes antérieurs d'estimation sur petits domaines (par exemple Elazar 2004 ; Chattopadhyay, Lahiri, Larsen et Reimnitz, 1999). Chen (2001) a proposé une modification hiérarchique au niveau de l'unité pour traiter le problème. De surcroît, certaines études (Cohen 2000) s'appuient sur la transformation logarithmique des estimations directes de la moyenne (ou du total) des données de comptage afin d'adopter un modèle d'estimation sur



# Utilisation de modèles multivariés pour l'estimation sur petits domaines du nombre de recrues dans les entreprises

Maria Rosaria Ferrante et Carlo Trivisano<sup>1</sup>

## Résumé

Le nombre de recrues dans les entreprises des zones locales de travail est un important indicateur de la réorganisation des processus de production locaux. En Italie, ce paramètre peut être estimé au moyen des données de l'Enquête Excelsior, bien que celle-ci ne fournisse pas d'estimations fiables pour les domaines d'intérêt. Dans le présent article, nous proposons une méthode d'estimation sur petits domaines multivariée appliquée à des données de comptage et basée sur la loi multivariée Poisson-Log-normale. Cette méthode servira à estimer le nombre de personnes recrutées par les entreprises pour remplacer les employés qui quittent ainsi que pour doter de nouveaux postes. Dans le cadre de l'estimation sur petits domaines, on suppose habituellement que les variances et les covariances d'échantillonnage sont connues. Cependant, ces dernières, de même que les estimations ponctuelles directes, sont instables. Étant donné la rareté du phénomène que nous analysons, les dénombrements dans certains domaines sont nuls, ce qui produit des estimations nulles des covariances des erreurs d'échantillonnage. Afin de tenir compte de la variabilité supplémentaire due à la matrice de covariances d'échantillonnage estimée et de résoudre le problème des variances et covariances insensées dans certains domaines, nous proposons une approche « intégrée » suivant laquelle nous modélisons conjointement les paramètres d'intérêt et les matrices de covariances des erreurs d'échantillonnage. Nous suggérons une solution de nouveau fondée sur la loi Poisson-Log-normale pour lisser les variances et les covariances. Les résultats que nous obtenons sont encourageants : le modèle d'estimation sur petits domaines fondé sur la loi multivariée normale-normale (MNN) et il rend possible une augmentation non négligeable de l'efficacité.

Mots clés : Loi multivariée Poisson-Log-normale ; dénombrements nuls ; fonction de variance généralisée ; modèles hiérarchiques bayésiens.

## 1. Introduction

Le nombre de personnes recrutées par les entreprises pendant une période déterminée peut être considéré comme un indicateur clé des changements en cours dans le système économique. Afin de mettre en relief la dynamique de la demande de main-d'œuvre locale, nous considérons le nombre de personnes recrutées par les entreprises dans les zones locales de marché de travail (LLMA pour *Local Labour Market Areas*), ces dernières étant regroupées selon (i) la spécialisation de production, (ii) la catégorie de taille des entreprises et (iii) le secteur industriel. Les domaines sont définis par recoupement de ces trois variables. Afin de mettre en relief les signes de réorganisation du processus de production, nous nous concentrons sur le nombre de « recrues remplaçant des employés quittant l'entreprise (recrues de substitution – RS) » et de « recrues occupant de nouveaux postes (nouvelles recrues – NR) ». En Italie, l'information au sujet des personnes recrutées par les entreprises est recueillie dans le cadre de l'Enquête Excelsior copartainée par l'Union des chambres de commerce italiennes (UNIONCAMERE), le ministère du Travail et l'Union européenne. Malheureusement, cette enquête ne fournit pas d'estimations fiables du nombre de personnes recrutées par les entreprises pour tous les domaines à cause de la petite

taille d'échantillon pour les petits domaines. Par conséquent, une méthode d'estimation sur petits domaines (EPD) a été adoptée afin d'obtenir des estimations dont le degré de variabilité est acceptable.

Dans le présent article, nous proposons une approche d'estimation sur petits domaines pour l'estimation de dénombrements. En raison de contraintes liées aux données, nous adoptons un modèle agrégé au niveau du domaine.

Puisque nous cherchons à estimer les nombres de RS et de NR, nous adoptons un modèle d'estimation sur petits domaines multivariée qui emprunte de l'information non seulement aux domaines, mais aussi aux corrélations entre les valeurs réelles des nombres de NR et de RS. Afin d'estimer le revenu médian de groupes de familles de diverses tailles, Fay (1987) a proposé un modèle de régression multivariée dans un contexte bayésien empirique. Des approches multivariées d'estimation sur petits domaines ont également été élaborées par Ghosh, Nangia et Kim (1996), par Datta, Fay et Ghosh (1991), par Datta, Ghosh, Nangia et Natarajan (1996) et par Datta, Lahiri, Maity et Lu (1999) pour des données continues dans le cadre des modèles hiérarchiques transversaux et chronologiques. Fabrizi, Ferrante et Paci (2005, 2008) ont adopté des modèles multivariés au niveau du domaine pour estimer un vecteur de paramètres continus de la pauvreté. Comme dans

1. Maria Rosaria Ferrante, Département de statistique, Université de Bologne, Italie. Courriel : maria.ferrante@unibo.it ; Carlo Trivisano, Département de statistique, Université de Bologne, Italie. Courriel : carlo.trivisano@unibo.it.



Cet argument reste applicable quand  $V_b$  est remplacée par son estimateur  $\hat{V}_b$  dans lequel sont utilisées des estimations à la place de  $\sigma_e^2$  et  $\sigma_\mu^2$ .

## Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Chambers, R. (2006). What is poverty? Who asks? Who answers? *Poverty in Focus*, UNDP, 3 et 4 décembre 2006.

Elbers, C., Lanjouw, J. et Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71, 355-364.

Elbers, C., Lanjouw, J. et Lanjouw, P. (2002). *Micro-level Estimation of Welfare*. Document de travail de recherche 2911, World Bank, Development Research Group, Washington, D.C.

Ghosh, M., et Rao J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Haslett, S., et Jones, G. (2004). *Local Estimation of Poverty and Malnutrition in Bangladesh*, Bangladesh Bureau of Statistics et United Nations World Food Programme.

Haslett, S., et Jones, G. (2005). *Local Estimation of Poverty in the Philippines: Philippine National Statistics Co-ordination Board/World Bank Report*. [http://siteresources.worldbank.org/INT/PGI/Resources/3426741092157888460/Local\\_Estimation\\_of\\_Poverty\\_Philippines.pdf](http://siteresources.worldbank.org/INT/PGI/Resources/3426741092157888460/Local_Estimation_of_Poverty_Philippines.pdf).

Haslett, S., et Jones, G. (2005). *Small Area Estimation of Poverty, Caloric Intake and Malnutrition in Nepal*. Published: Nepal Central Bureau of Statistics/World Food Programme, United Nations/World Bank, septembre 2006, 184pp, ISBN 999337018-5.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226-252.

Horton, N.J., et Lipsitz, S.R. (1999). Review of software to fit Generalized Estimating Equation regression models. *The American Statistician*, 53, 160-169.

Liang, K.L., et Zeger, S. (1986). Longitudinal data analysis using Generalized Linear Models. *Biometrika*, 73, 13-22.

Millitino, A.F., Ugarte, M.D., Goicoa, T. et Gonzalez-Audiciana, M. (2006). Using small area models to estimate the total area occupied by olive trees. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 450-461.

ONU site web. <http://www.un.org/f/m/millenniumgoals/>.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B*, 60, 23-40.

Pfeffermann, D., Moura, F.A. et Silva, P.L. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 949-959.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Rao, J.N.K. (1999). Quelques progrès récents concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquête*, 25, 199-212.

Rao, J.N.K. (2003). *Small Area Estimation*, Wiley Series in Survey Methodology. Wiley-Interscience, John Wiley & Sons, Inc.

NSCB (2000). *Profile of Censuses and Surveys*. National Statistical Coordination Board, Philippines.

Stiegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Series in Psychology. New York : McGraw-Hill.

Skinner, C.J., Holt, D. et Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Chichester : John Wiley & Sons.

You, Y., et Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *La Revue Canadienne de Statistique*, 30, 431-439.

You, Y., Rao, J.N.K. et Kovacevic, M. (2003). Estimation des effets fixes et des composantes de la variance par un modèle à valeur aléatoire à l'origine en utilisant des données d'enquête. *Recueil : Symposium 2003, Défis reliés à la réalisation d'enquêtes pour la prochaine décennie*. Statistique Canada.

Zhao, Q. (2006). User manual for PovMap, The World Bank. [http://siteresources.worldbank.org/INT/PGI/Resources/3426741092157888460/Zhao\\_ManualPovMap.pdf](http://siteresources.worldbank.org/INT/PGI/Resources/3426741092157888460/Zhao_ManualPovMap.pdf).

elle ne repose pas sur un fondement théorique solide. Nous recommandons de remplacer la partie de la méthode ELL correspondant à l'ajustement d'un modèle aux données d'enquête. Les autres méthodes prises en considération (pseudo-MPLSB, IEFP et RGE) ont toutes un fondement mathématique théorique valide et les résultats produits peuvent être interprétés clairement une fois que les hypothèses ont été vérifiées. Appliquées à des données d'enquête complexes pondérées recueillies aux Philippines, les diverses méthodes indiquent que, pour l'estimation des composantes de la variance d'après des données d'enquête et, donc, pour l'estimation sur petits domaines à un fin niveau de détail, la méthode du pseudo-MPLSB et particulièrement la méthode IEFP donnent vraisemblablement de meilleurs résultats que les méthodes RGE ou ELL, quoique la méthode RGE est valable et facile à utiliser parce qu'elle est disponible dans les logiciels du commerce.

Nous avons également montré qu'au niveau où l'estimation sur petits domaines est effectivement utilisée pour répartir l'aide, l'estimation de la variance des estimations sur petits domaines a tendance à être dominée par la variance au niveau de la grappe plutôt que par l'exactitude des estimations des paramètres de régression. Donc, il est particulièrement important que la composante de la variance au niveau de la grappe (et, si l'ajustement du modèle est exécuté tel que recommandé, toute composante de la variance au niveau du petit domaine) soit estimée correctement. Il importe aussi que le modèle de régression utilisé pour produire les estimations sur petits domaines (y compris le choix de variables indépendantes pertinentes) soit approprié. Essentiellement, aux niveaux plus faibles d'agrégation, l'erreur-type des estimations sur petits domaines est dominée par les composantes de la variance, de sorte que l'estimation de ces dernières est cruciale, quel que soit le choix du niveau d'agrégation. Une méthode de régression sur données d'enquête valable, le bon choix des variables de régression et la détermination minutieuse de la taille d'échantillon (surtout si des modèles de régression distincts sont ajustés à des sous-ensembles de données d'enquête) demeurent aussi des éléments essentiels à une bonne estimation sur petits domaines de la pauvreté dans le tiers monde.

## Remerciements

Les auteurs remercient les examinateurs et le rédacteur associé de leur lecture attentive du manuscrit et de leurs suggestions utiles.

## Annexe

Dans la note en bas de page 8 du document de travail de la Banque mondiale rédigé par Elbers et coll. (2002) et, implicitement, dans l'article de Elbers et coll. (2003) publié dans *Econometrica*, la covariance du processus d'erreur est désignée par  $\Omega$  et il est déclaré que  $W\Omega^{-1} = P^T P$ , où  $W$  est « une matrice de pondération de facteurs d'extension ». Dans la notation de la section 4 du présent article,  $W$  est diagonale par bloc, ou diagonale, avec blocs diagonaux  $V_b$ , et  $\Omega$  est diagonale par bloc avec blocs diagonaux  $V_b$ . Cependant, soit  $W$  et  $\Omega$  (ou  $\Omega^{-1}$ ) sont non conformables (avec les facteurs de pondération dans  $W$  au niveau de la grappe et les observations, et donc  $\Omega^{-1}$ , au niveau individuel), soit, si elles sont conformables,  $W\Omega^{-1}$  est généralement asymétrique (même si  $W$  est diagonale) à moins que  $W$  soit un multiple simple de la matrice identité, c'est-à-dire  $W = \sigma^2 I$ .

D'où,  $W\Omega^{-1}$  n'est pas égale à  $P^T P$  comme il l'a été soutenu, puisque  $P^T P$  est symétrique en général et que  $W\Omega^{-1}$  ne l'est pas. Rendre  $W\Omega^{-1}$  symétrique en l'additionnant à sa transposée et en divisant le résultat par deux, comme cela est fait dans le logiciel PovMap de la Banque mondiale, n'est pas une solution techniquement adéquate de ce problème. (Notons que même dans le cas simple où  $W$  et  $\Omega^{-1}$  sont conformables, et que  $W$  est diagonale, mais que tous les éléments diagonaux ne sont pas égaux,  $W\Omega^{-1}$  n'est pas diagonale, parce que chaque élément de la ligne  $i$  de  $\Omega^{-1}$  est multiplié par  $w_i$  (où  $w_i$  est le  $i^{\text{e}}$  élément diagonal de  $W$ ), mais que, dans la  $i^{\text{e}}$  colonne, chaque élément  $n$  est pas multiplié par un poids identique.)

En mettant de côté cette question de symétrie et en utilisant  $P^T P$  à la place de  $W\Omega^{-1}$ , Elbers et coll. semblent affirmer que comparer leur « estimateur MCG pondéré corrigé pour les données d'enquête » à l'estimateur MCG non corrigé implique qu'au lieu d'utiliser  $\Omega^{-1}$  comme mesure sous-jacente (c'est-à-dire l'inverse de la matrice de covariance pertinente), une version pondérée, à savoir  $W\Omega^{-1} W^T$ , devrait être utilisée. Cela ne crée aucun problème d'asymétrie en soi (à condition d'utiliser  $P^T P$  à la place de  $W\Omega^{-1}$ ). Toutefois, même si  $W$  était diagonale et que  $P^T P$  était utilisé, la matrice de pondération  $W$  ne peut même pas utiliser les poids diagonaux inégaux correspondant aux unités échantillonnées, disons  $w_i$ , parce que le  $i^{\text{e}}$  élément de  $\Omega^{-1}$  (contrairement au  $i^{\text{e}}$  élément de  $\Omega$ ) ne correspond pas aux  $i^{\text{e}}$  et  $j^{\text{e}}$  unités dans l'échantillon (ou dans la population), de sorte que l'on ne sait pas vraiment ce qu'est  $W$  ou comment  $W$  peut être définie de manière sensée comme une « matrice de pondération de facteurs d'extension ».



Tableau 4

Test de Kruskal-Wallis pour les variances estimées au niveau municipal (N = 1 243)

Méthodes		Effet de grappe		Effet bêta		Variance	
d'EPD	Médiane	Rang moyen	Z	Médiane	Rang moyen	Z	Rang moyen
	ELL (sans hétérosc.)	2 961,2(a)	-3,22	0,0002311	3 067,3(ab)	-0,89	2 963,4(a)
ELL (avec hétérosc.)	0,002843	2 961,2(a)	-3,22	0,0002128	2 802,0(c)	-6,72	2 930,8(a)
Pseudo-MPLSB	0,003094	3 229,4(b)	2,67	0,0002449	3 257,5(ad)	3,28	3 241,3(b)
IEEP	0,003294	3 426,9(b)	7,01	0,0002529	3 364,5(d)	5,64	3 441,3(b)
RGE (Stata)	0,002843	2 961,2(a)	-3,22	0,0002311	3 048,7(b)	-1,3	2 963,1(a)
Global	3,108	3,108		3,108	3,108		3,108
Statistique de KW		H = 69,92	(P = 0,000)	H = 72,19	(P = 0,000)	H = 78,06	(P = 0,000)

Tableau 5

Test de Kruskal-Wallis pour les variances estimées au niveau provincial (N = 83)

Méthodes		Effet de grappe		Effet bêta		Variance	
d'EPD	Médiane	Rang moyen	Z	Médiane	Rang moyen	Z	Rang moyen
	ELL (sans hétérosc.)	0,0002518	-0,65	0,0001162	207,7	-0,03	0,00039
ELL (avec hétérosc.)	0,0002518	200,3	-0,65	0,0001095	190,1	-1,52	0,00038
Pseudo-MPLSB	0,000274	214,9	0,59	0,0001239	224,2	1,37	0,00042
IEEP	0,0002916	224,2	1,38	0,0001287	224,1	2,22	0,00045
RGE (Stata)	0,0002517	200,3	-0,65	0,00010	184	-2,04	0,00037
Global	208	208		208	208		208
Statistique de KW		H = 2,82	(P = 0,589)	H = 10,61	(P = 0,031)	H = 4,48	(P = 0,344)

Tableau 6

Test de Kruskal-Wallis pour les variances estimées au niveau régional (N = 16)

Méthodes		Effet de grappe		Effet bêta		Variance	
d'EPD	Médiane	Rang moyen	Z	Médiane	Rang moyen	Z	Rang moyen
	ELL (sans hétérosc.)	0,000050	-0,45	0,000077	40,9	0,08	0,00013
ELL (avec hétérosc.)	0,000050	38,2	-0,45	0,000073	35,1	-1,05	0,00012
Pseudo-MPLSB	0,000055	42,6	0,4	0,000082	46,9	1,23	0,00014
IEEP	0,000058	45,3	0,93	0,000085	50,1	1,85	0,00015
RGE (Stata)	0,000050	38,2	-0,45	0,000070	29,6	-2,1	0,00013
Global	40,5	40,5		40,5	40,5		40,5
Statistique de KW		H = 1,30	(P = 0,861)	H = 8,36	(P = 0,079)	H = 2,58	(P = 0,630)

Naturellement, ce genre de considérations (quoique essentielles) doivent s'appuyer sur une épuraton appropriée des données, une bonne concordance des variables indépendantes possibles (en ce qui concerne la moyenne, la variance et la signification) de l'enquête et du recensement, dans le cas où des données de recensement sont également utilisées. Sont également nécessaires l'examen approprié, qui prend beaucoup de temps, d'une grande gamme de variables indépendantes possibles et la reconnaissance des limites qu'imposent les petites tailles d'échantillon à la subdivision des données d'enquête, puisque toutes les erreurs-types estimées pour les estimations des paramètres de régression ainsi que les estimations sur petits domaines (quelle que soit la méthode utilisée pour ajuster l'estimation des composantes de la variance) reposent sur la condition que le modèle de régression est correct.

## 8. Conclusion et recommandation

De bonnes statistiques sur la pauvreté sont nécessaires afin de pouvoir surveiller efficacement les interventions et

L'aide offerte aux diverses localités appauvries. Les méthodes d'estimation sur petits domaines constituent l'une des méthodologies utilisées pour produire ce genre de statistiques. En ce sens, les questions soulevées dans le présent article au sujet de l'exactitude des estimations sur petits domaines ne sont pas simplement théoriques, mais sont essentielles à la réalisation des Objectifs du millénaire pour le développement et à la répartition de l'aide dans un secteur d'activités où les enjeux se chiffrent à plusieurs milliards de dollars.

Dans le présent article, nous avons examiné quatre méthodes d'estimation pour ajuster les modèles de régression en utilisant des données d'enquête et nous les avons reliées à l'estimation de la pauvreté sur petits domaines. Nous avons montré que, même si les écarts entre les estimations sont insuffisants pour rendre invalides les études à l'échelle nationale publiées, la méthode d'ajustement d'un modèle à des données d'enquête la plus fréquemment mise en œuvre, c'est-à-dire la méthode ELL avec hétéroscé-dasticité recommandée par la Banque mondiale, présente certaines limites, car, comme sa version homoscédaistique,



tendance à être semblables. Explicitement, l'examen du tableau 4 montre que le classement des estimations de la variance concorde généralement avec le classement des effets de grappe.

Dans l'estimation de la pauvreté, les estimations aux

niveaux élevés d'agrégation, telles que celles des tableaux 5 et 6, sont généralement produites en vue de les comparer aux estimations directes sur données d'enquête à ces niveaux d'agrégation plus élevés, même si elles ne sont pas particulièrement utiles pour la répartition de l'aide. Néanmoins, les résultats corroborent ceux présentés pour le niveau plus faible d'agrégation. Aux tableaux 5 et 6, les variances estimées des estimations de la pauvreté produites par les diverses méthodes ne diffèrent pas significativement les unes des autres aux niveaux provincial et régional, effet attribuable en partie au petit nombre de provinces et au nombre encore plus petit de régions. Les variances, et donc les erreurs-types, ne sont peut-être pas significativement différentes les unes des autres, mais il faut souligner que la

méthode RGF a tendance à produire l'estimation la plus faible de l'erreur-type pour le modèle de régression et, à son tour, l'estimation la plus faible de la variance pour la pauvreté au niveau régional, même si cette méthode produit des erreurs-types plus élevées pour les coefficients de régression individuels (qui correspondent aux éléments diagonaux uniquement dans la matrice de covariance estimée de  $\beta$ ). Comme prévu, à un niveau encore plus élevé d'agrégation, pour toutes les méthodes, l'effet relatif de la composition de la régression est encore plus prononcé.

La conclusion générale est, que l'on ajuste un modèle à des données d'enquête uniquement ou que l'on utilise les estimations des paramètres de régression sur données d'enquête conjuguées à des données de recensement, qu'il est essentiel non seulement de trouver un modèle (c'est-à-dire, un ensemble de variables indépendantes) approprié fondé sur une taille d'échantillon adéquate, mais aussi d'obtenir de bonnes estimations des paramètres de régression et de leurs erreurs-types sous ce modèle, ainsi que de bonnes estimations et composantes de la variance à tous les niveaux pertinents d'agrégation. Habituellement, les niveaux pertinents d'agrégation sont déterminés par le plan de sondage, plutôt que simplement d'après le niveau auquel les estimations sur petits domaines sont souhaitées, quoique le nombre de niveaux ne doive pas nécessairement être limité à deux (par exemple niveau de la grappe et niveau du ménage).

(Qu'elles soient utilisées pour l'estimation de la pauvreté ou dans un autre contexte, les données d'enquête intro-

duisent aussi des problèmes ayant trait aux poids de sondage

qui peuvent être importants non seulement pour l'estimation des paramètres de régression (et de leurs erreurs-types), mais aussi pour l'estimation des composantes de la variance. L'intégration des poids de sondage dans les modèles de régression avec données corrélées pose des problèmes, parce que c'est la corrélation de population, telle qu'elle est appliquée aux données d'enquête pondérées qui doit être modélisée correctement, de sorte que pondérer les matrices de corrélation par multiplication de matrices (comme cela est fait dans la méthode ELL) n'est pas une technique

appropriée (voir l'annexe).

Pour les données sur les Philippines et pour la liste spécifiée de variables indépendantes, quelle que soit la méthode utilisée parmi les quatre étudiées, les estimations des paramètres sont très semblables, ce qui donne à penser qu'une question plus importante est la sous-estimation possible des erreurs-types des estimations des paramètres et des composantes de la variance, particulièrement au niveau de la grappe. La méthode ELL est la moins prudente en ce sens qu'elle donne les estimations les plus faibles des deux mesures de variance, et à cet égard (comme en ce qui concerne son utilisation de matrices de covariance estimées asymétriques) une certaine mise en garde pourrait être justifiée en ce qui concerne les aspects régression et composition de la variance de cette méthode. La méthode RGF produit des estimations des erreurs-types des estimations sur petits domaines semblables à celles de la méthode ELL lorsque l'on utilise la même méthode pour les composantes de la variance, bien qu'elle donne des erreurs-types plus grandes (et utilise une bonne matrice de covariance) pour les paramètres de régression. Il en est ainsi parce qu'à un niveau moins aggré, c'est-à-dire le niveau auquel la plupart des estimations sur petits domaines sont effectivement utilisées, les composantes de la variance dominent.

Dans les méthodes pseudo-MPLSB et IIEEP, les poids de sondage sont intégrés correctement (étant donné un choix approprié de la pseudo-vraisemblance et, donc, des EEG) et les estimations de la composante de la variance au niveau de la grappe sont plus grandes (c'est-à-dire plus prudentes). Cela donne à penser que ces deux méthodes, particulièrement la méthode IIEEP, comptent parmi les meilleures disponibles à l'heure actuelle, pas nécessairement pour l'estimation des équations de régression (pour laquelle l'existence de logiciels standard pourrait donner un avantage à la méthode RGF), mais pour estimer les composantes cruciales de la variance.

L'effet de la variation au niveau de la grappe est différent : aux niveaux plus faibles d'agrégation (par exemple, municipalité), la variance calculée des estimations sur petits domaines est dominée par la composante au niveau de la grappe ou effet au niveau de la grappe, ce qui signifie que, pour l'estimation sur petits domaines (autres que les estimations régionales), la composante de la variance et non le modèle de régression a l'effet le plus important sur la valeur de l'erreur-type des estimations sur petits domaines. Par conséquent, l'exactitude des estimations des composantes de la variance, surtout au niveau de la grappe, est un élément essentiel à l'estimation exacte de l'erreur-type des estimations sur petits domaines au niveau d'agrégation auquel elles sont le plus utiles (par exemple au niveau municipal aux Philippines). De nouveau, la méthode utilisée pour l'ajustement de la phase 1 pour les composantes de la variance dont nous avons discuté dans le présent article joue un rôle crucial dans l'estimation sur petits domaines de la pauvreté.

Aux tableaux 4 à 6, nous présentons les résultats du test de Kruskal-Wallis (Siegel (1956) pour les diverses méthodes d'ajustement effectué sur les variances estimées au niveau municipal (tableau 4), provincial (tableau 5) et régional (tableau 6). Au tableau 4, des écarts significatifs existent entre les estimations de la variance produites par les diverses méthodes d'estimation sur petits domaines, comme l'indiquent les valeurs  $p$  de la statistique de Kruskal-Wallis (KW). La comparaison multiple des rangs moyens montre que les méthodes du pseudo-MPLSB et IEFP donnent des estimations de la variance au niveau de la grappe qui sont significativement plus élevées que celles obtenues pour les autres méthodes, mais qu'elles ne diffèrent pas significativement l'une de l'autre (quoique, pour la méthode IEFP, la valeur  $Z$  pour l'écart par rapport au rang moyen est en général assez bien plus élevée que toutes les autres).

La méthode ELL et la méthode RGE produisent des estimations des composantes de la variance significative-ment plus faibles et similaires. Ce résultat tient principalement au fait que nous avons utilisé la méthode ELL d'estimation des composantes de la variance pour la méthode RGE (parce que cette dernière ne comporte habituellement pas l'estimation de ces composantes), quoique les résidus que nous avons utilisés n'étaient pas identiques pour les deux méthodes d'ajustement de la régression. Comme prévu, au niveau municipal pour lequel les estimations sur petits domaines ont été utilisées en pratique, l'effet de grappe (plutôt que l'incertitude associée aux coefficients de régression) est généralement la partie dominante dans les estimations de la variance des estimations sur petits domaines. Puisque la variance au niveau de la grappe est la même pour les méthodes ELL et RGE, les estimations correspondantes de la variance au niveau du petit domaine ont également

méthode ELL. En calculant de la même manière la moyenne de (7) pour obtenir la moyenne réelle  $\bar{Y}_a$ , en soustrayant le résultat de (26) et en appliquant l'opérateur de variance, nous obtenons l'équation de la variance de l'erreur de prédiction :

$$V(\bar{y}^a - \bar{Y}^a) = \mathbf{X}_a^a \Phi^w \mathbf{X}_a^a + \frac{1}{\sum_{b=1}^a N_b^a} N_b^a \sigma_v^2 + \frac{1}{N_a^a} \sigma_e^2 \quad (27)$$

où  $N_a^a$  est la taille de la population à un niveau particulier d'agrégation,  $N_b^a$  est la taille de la population dans chaque grappe,  $\Phi^w$  est la matrice de variance-covariance des estimations des coefficients de régression, et  $(\sigma_v^2, \sigma_e^2)$  sont les composantes de la variance au niveau de la grappe et du ménage, respectivement. Notons qu'estimer cette variance de l'erreur de prédiction requiert des estimations des composantes de la variance, mais que tout biais causé par l'incertitude dans ces dernières serait un effet de deuxième ordre (voir Prasad et Rao 1990).

Selon l'expression (27), l'importance de l'influence exercée par le modèle de régression fondé sur les données d'enquête et les autres composantes de la variance (au niveau de la grappe et du ménage) sur l'exactitude des estimations sur petits domaines finales peut être comparée pour toute méthode d'ajustement et (ou) tout niveau d'agrégation. En général, le modèle de régression (par la voie de l'estimation des paramètres de régression) ou l'effet de grappe est le facteur qui domine l'exactitude estimée de l'estimation sur petits domaines calculée. En utilisant le modèle au niveau national, dont les données sont présentées au tableau 1, et les variables auxiliaires de l'enquête, au lieu du recensement, pour estimer le premier terme de (27), nous constatons que la mesure dans laquelle l'effet du modèle de régression contribue à la variance des estimations sur petits domaines augmente appréciablement, à mesure que les données sur les ménages sont plus agrégées – environ 0,25 % au niveau municipal, 20 % au niveau provincial et 70 % au niveau régional. Autrement dit, la dominance de l'incertitude de l'estimation des paramètres du modèle de régression est d'autant plus importante que les données sont agrégées dans de plus grands domaines, indépendamment de la méthode d'ajustement de la régression. Ce résultat est conforme aux attentes, car même à des niveaux élevés d'agrégation, la contribution de l'effet du modèle à la variance globale dépend des valeurs moyennes des covariables, et non de la taille de la population. C'est pour cette raison qu'au niveau régional le plus agrégé, les méthodes d'estimation sur petits domaines offrent habituellement peu d'amélioration par rapport aux estimations directes. C'est également pour cela qu'il est important (comme nous l'avons fait dans le présent article) d'examiner en détail les procédures d'ajustement des régressions appliquées dans l'estimation sur petits domaines de la pauvreté dans le tiers monde.



de régression aux données d'enquête, à la deuxième phase, ce modèle est appliqué aux données de recensement en tant que prédicteur au niveau du ménage, autrement dit l'équation de régression (quelle que soit la façon dont elle a été estimée) est utilisée pour trouver les valeurs prévues du revenu et des dépenses par tête pour chaque ménage du recensement, produites au moyen de

$$Y_{bh} = \mathbf{x}'_{bh} \boldsymbol{\beta} + v_b + e_{bh} \quad (25)$$

en utilisant des valeurs imputées de  $v_b$  et  $e_{bh}$  (fondées, par exemple, sur un échantillonnage bootstrap de leurs estimations sur données d'enquête). Ici, les  $\mathbf{x}_{bh}$  sont les variables auxiliaires provenant du recensement. Les indices de pauvreté sont habituellement fondés sur des fonctions non linéaires du logarithme du revenu ou du logarithme des dépenses, si bien que les prédictions issues de (25) sont transformées comme il convient avant de calculer la moyenne sur chaque petit domaine. Notons qu'en pratique,  $v_b$  peut être estimé pour les grappes échantillonnées, mais les codes d'échantillon et de recensement ne concordent habituellement pas, de sorte que celles-ci ne peuvent pas être identifiées dans le recensement et c'est donc le bootstrap (en sélectionnant parmi les barangays échantillonnés, c'est-à-dire les LPE) qui fournit les valeurs imputées pour tous les barangays; un commentaire parallèle s'applique aux  $e_{bh}$  pour les ménages à l'intérieur des grappes. L'avantage général de l'utilisation de données de recensement de cette façon (comme le fait la méthode ELL) est que les variables prédictives peuvent être utilisées pour tous les ménages du recensement (qui sont nombreux) plutôt que simplement ceux de l'enquête, ce qui augmente la précision des estimations sur petits domaines (à condition que le modèle soit correct). Notons que les estimations données par (25) demeurent sans biais, même si  $v_b$  et  $e_{bh}$  ne sont pas inclus dans la prédiction proprement dites; mais les estimations de la variance pour le petit domaine  $a$  doivent être calculées en se basant sur l'équation (25) afin que soit intégrée la variance supplémentaire nécessaire au niveau de la grappe et du ménage.

Dans l'estimation de la pauvreté, nous nous intéressons à des statistiques sommées au niveau du domaine pour des fonctions non linéaires de  $Y_{bh}$ , comme savoir si l'estimation est inférieure au seuil de pauvreté (prévalence de la pauvreté) et quel est l'écart de pauvreté, plutôt qu'à l'ajustement de la régression proprement dit. Il est intéressant, ici, d'examiner les effets de l'incertitude du modèle sur les estimations de la moyenne de domaine.

$$\bar{y}_a = \mathbf{x}'_a \boldsymbol{\beta} \quad (26)$$

où  $\mathbf{x}_a$  est la moyenne de population (c'est-à-dire du recensement) pour le domaine  $a$  des covariables incluant la constante 1, après avoir appliqué le modèle de régression aux données de recensement comme à la phase 2 de la

variation au niveau de la grappe dans (7) qui doit être prise en considération (à divers degrés selon le niveau d'aggrégation utilisé pour construire les petits domaines), ainsi qu'une variation au niveau du ménage. Ces sources supplémentaires de variation peuvent être évaluées par l'estimation des composantes de la variance. Comme il est montré ci-dessus, indépendamment du niveau (national, régional ou provincial) auquel le modèle est formulé, la méthode IIEP produit la variance au niveau du ménage la plus faible, tandis que la méthode ELL produit la variance au niveau de la grappe la plus faible. Puisque la variation au niveau de la grappe contribue habituellement plus fortement à l'erreur-type estimée au niveau du petit domaine, la méthode ELL est de nouveau la moins prudente. Nous notons que la variance au niveau du ménage sous la méthode ELL avec modèle d'hétéroscédasticité varie d'une unité à l'autre, si bien que nous présentons la valeur moyenne, et que le  $R^2$  estimé pour le modèle d'hétéroscédasticité est négligeable.  $R^2 = 0,03$ , même au niveau national, de sorte qu'en ce qui concerne au moins l'ajustement du modèle de régression, cette méthode peut offrir quelques avantages pour l'ensemble de données examinées ici. Notre expérience de l'application de la méthode ELL nous porte à conclure que la modélisation de l'hétéroscédasticité n'est pas nécessaire.

Si nous revenons à la régression (c'est-à-dire les estimations produites pour  $\boldsymbol{\beta}$  et l'erreur-type estimée pour les diverses méthodes), la méthode IIEP est celle qui intègre le mieux les poids de sondage provenant du calcul des composantes de la variance nécessaires pour produire les estimations sur petits domaines et leurs erreurs-types estimées. En ce qui concerne l'exécution, la méthode RGE serait généralement l'option la plus simple, parce qu'elle est disponible, par exemple, dans des logiciels tels que Stata, Sudaan ou WesVar. La méthode ELL combine les poids de sondage et la structure de covariance d'une façon non standard, en ce sens qu'elle utilise une estimation de  $\mathbf{W}^b \mathbf{V}^{b-1}$  dans (8) et (9) pour produire une matrice de covariance estimée asymétrique pour les estimations de  $\boldsymbol{\beta}$  et pour estimer  $\boldsymbol{\beta}$  proprement dit. Dans ce dernier cas, cette estimation serait acceptable si la matrice asymétrique était une inverse généralisée de la matrice de covariance correcte. Cependant, elle n'est manifestement pas acceptable comme matrice de covariance estimée, problème que la méthode ELL essaye de contourner (par exemple dans le logiciel POVMAP de la Banque mondiale) par calcul de la moyenne de chacune des paires pertinentes d'éléments hors diagonale afin de satisfaire la condition nécessaire qu'une matrice de covariance soit symétrique.

En général, dans la méthode ELL d'estimation de la pauvreté, seules peuvent être utilisées les variables dont la moyenne et l'écart-type concordent dans les moyennes sur données d'enquête ainsi que sur données de recensement. Cette contrainte tient au fait qu'après avoir ajusté le modèle



**Tableau 2**  
Estimations au niveau régional des paramètres de régression avec les erreurs-types et les composantes de la variance pour les quatre méthodes. \*Valeur différente pour chaque ménage (moyenne = 0,18930) \*\*Basé sur les résultats pour la méthode ELL

Variables	ELL(sans hétérosc.)		ELL(avec hétérosc.)		Pseudo-MPLSB		IEEP		RGE	
	Bêta	Erreur-type	Bêta	Erreur-type	Bêta	explicatives	Bêta	Erreur-type	Bêta	Erreur-type
famsizescg	-0,12227	0,00760	-0,12934	0,00689	-0,12377	0,00752	-0,12380	0,00749	-0,11786	0,00997
famsizescg	0,01096	0,00164	0,01190	0,00147	0,01101	0,00163	0,01102	0,00162	0,01030	0,00195
dom_help	0,81037	0,08873	0,75624	0,10986	0,80727	0,08784	0,80708	0,08751	0,84490	0,08911
wall_highr	-0,06808	0,04289	-0,06390	0,04743	-0,06020	0,04272	-0,05973	0,04257	-0,14472	0,04226
wall_strong	0,13761	0,03745	0,15212	0,03469	0,14514	0,03737	0,14560	0,03725	0,06116	0,04249
fa_xs	-0,22074	0,04910	-0,22268	0,04518	-0,22723	0,04875	-0,22761	0,04858	-0,14856	0,05665
fa_s	-0,13540	0,03840	-0,12255	0,03344	-0,13775	0,03805	-0,13789	0,03791	-0,11059	0,04538
fa_l	0,09484	0,03709	0,08894	0,03429	0,09590	0,03676	0,09597	0,03663	0,08529	0,04122
fa_xl	0,16627	0,04315	0,15519	0,04072	0,16938	0,04284	0,16958	0,04269	0,13698	0,04897
fa_xxl	0,33706	0,04545	0,31196	0,04829	0,34173	0,04516	0,34201	0,04500	0,29156	0,05148
fa_xxxl	0,33103	0,06185	0,30377	0,06029	0,33762	0,06134	0,33801	0,06111	0,26052	0,06635
all_hsed	0,33987	0,05253	0,35591	0,04783	0,33807	0,05209	0,33796	0,05189	0,35776	0,04843
all_coed	1,21824	0,05734	1,24762	0,05842	1,20787	0,05692	1,20726	0,05671	1,32979	0,06227
per_kids	-0,24699	0,06440	-0,24047	0,05846	-0,24439	0,06371	-0,24424	0,06347	-0,27423	0,07050
per_61up	-0,14609	0,06126	-0,15938	0,05787	-0,14703	0,06063	-0,14708	0,06040	-0,13525	0,07124
hou_9600	1,13985	0,49103	1,27035	0,47888	1,14320	0,52137	1,14357	0,52172	1,07509	0,51937
Hou_own_ref	1,45233	0,24550	1,51020	0,23864	1,44986	0,26072	1,44985	0,26089	1,44779	0,23585
const	9,36877	0,20322	9,32363	0,19660	9,36597	0,21502	9,36569	0,21512	9,41385	0,21430
Estimation des composantes de la variance	Niveau du ménage	0,19544	Niveau du ménage	0,03073	Niveau du ménage	0,19052	Niveau du ménage	0,03728	Niveau du ménage	0,19544
	Niveau de la grappe	0,03073	Niveau de la grappe	NA*	Niveau de la grappe	0,03073	Niveau de la grappe	0,03728	Niveau de la grappe	0,03073

**Tableau 3**  
Estimations au niveau provincial des paramètres de régression avec les erreurs-types et les composantes de la variance pour les quatre méthodes. \*Valeur différente pour chaque ménage (moyenne = 0,23749) \*\*Basé sur les résultats pour la méthode ELL

Variables	ELL(sans hétérosc.)		ELL(avec hétérosc.)		Pseudo-MPLSB		IEEP		RGE	
	Bêta	Erreur-type	Bêta	Erreur-type	Bêta	Erreur-type	Bêta	Erreur-type	Bêta	Erreur-type
famsizescg	-0,1450	0,0175	-0,1489	0,0156	-0,1452	0,0179	-0,1449	0,0171	-0,1413	0,0097
famsizescg	0,0090	0,0067	0,0124	0,0065	0,0091	0,0065	0,0092	0,0062	0,0055	0,0055
fa_xs	-0,4549	0,1126	-0,3816	0,1010	-0,4552	0,1149	-0,4546	0,1095	-0,4479	0,0718
fa_s	-0,2550	0,0976	-0,2653	0,0794	-0,2545	0,0995	-0,2555	0,0951	-0,2693	0,1198
wall_highr	-0,2055	0,0945	-0,1474	0,0778	-0,2057	0,0965	-0,2058	0,0919	-0,2063	0,1070
all_hsed	0,4007	0,1643	0,3531	0,1448	0,4015	0,1673	0,4006	0,1601	0,3891	0,1585
all_coed	1,5411	0,1677	1,8202	0,1769	1,5429	0,1709	1,5429	0,1635	1,5439	0,2326
Hou_own_ref	3,4373	1,0270	3,2630	1,0582	3,4265	1,0622	3,4274	0,9871	3,4392	0,5733
per_wor_prh	-1,1075	1,1933	-1,5801	1,2008	-1,1049	1,2327	-1,1056	1,1483	-1,1150	0,8729
const	10,0976	0,1480	10,0798	0,1279	10,0988	0,1517	10,0981	0,1435	10,0872	0,1373
Estimation des composantes de la variance	Niveau du ménage	0,25753	Niveau du ménage	0,25753	Niveau du ménage	0,26682	Niveau du ménage	0,24498	Niveau du ménage	0,25753
	Niveau de la grappe	0,01871	Niveau de la grappe	NA*	Niveau de la grappe	0,02079	Niveau de la grappe	0,01671	Niveau de la grappe	0,01871

53 % et par rapport à la méthode IEEP, d'environ 48 %. Pour un certain nombre de provinces, la méthode IEEP a tendance à produire l'estimation la plus faible de la variance au niveau de la grappe. En ce qui concerne la variance du ménage, la méthode IEEP produit encore l'estimation la plus faible. En général, les estimations de la variance au niveau de la grappe ont tendance à être plus variables au niveau provincial, ce qui est dû aux plus petites tailles d'échantillon.

En ce qui concerne les estimations de la pauvreté sur petits domaines, après l'application du modèle de régression aux données de recensement, les erreurs-types estimées dans la régression représentent une partie seulement des erreurs-types des estimations sur petits domaines. Il existe aussi une

Comme pour les estimations au niveau régional, les estimations des coefficients de régression au niveau provincial sont similaires, à l'exception de certaines différences pour les estimations par les méthodes RGE et ELL<sub>H</sub>. Pour ce qui est des erreurs-types estimées des coefficients de régression, la méthode ELL<sub>H</sub> produit encore les estimations les plus faibles pour la majorité des coefficients des caractéristiques du ménage ; toutefois, la méthode RGE (au lieu de la méthode ELL<sub>H</sub>) produit maintenant les erreurs-types estimées les plus faibles pour la majorité des moyennes municipales. La méthode ELL continue d'avoir tendance à générer l'estimation la plus faible de la variance au niveau de la grappe pour la plupart des provinces, le ratio le plus petit par rapport au pseudo-MPLSB étant d'environ

pour toutes les moyennes municipales. La méthode ELL\_H (ELL avec hétéroscédasticité) peut être considérée comme étant la moins prudente, puisqu'elle produit les erreurs-types les plus faibles pour tous les coefficients de régression estimés des caractéristiques au niveau du ménage, ainsi que pour les moyennes municipales, sauf dans le cas de deux variables, pour lesquelles la méthode RGE a produit les estimations des composantes de la variance, la méthode ELL produit la variance estimée au niveau de la grappe la plus faible, correspondant à environ 92 % de celle obtenue par la méthode du pseudo-MPLSB et à 86 % de celle obtenue par la méthode IIEFP. Pour ce qui est de la variance au niveau du ménage, la méthode IIEFP est celle qui produit l'estimation la plus faible.

(ou) ELL\_H produisent, pour quelques variables, des estimations significativement différentes de celles obtenues par les autres méthodes. Comme dans le cas des erreurs-types estimées au niveau national, la méthode RGE a tendance à être la plus prudente pour la majorité des modèles de niveau régional, ayant donné les erreurs-types estimées les plus élevées pour la plupart des coefficients de régression des caractéristiques du ménage. La méthode IIEFP produit l'erreur-type estimée la plus grande pour la plupart des coefficients des moyennes municipales. La méthode ELL\_H produit les erreurs-types les plus faibles pour la majorité des coefficients de régression des caractéristiques du ménage et des moyennes municipales. La méthode ELL a tendance à produire l'estimation la plus faible de la variance au niveau de la grappe, les ratios par rapport au pseudo-MPLSB et à la méthode IIEFP variant d'environ 82 % à 100 %. La méthode IIEFP est encore celle dominant la variance au niveau du ménage la plus faible.

**Tableau 1**  
Estimations au niveau national des paramètres de régression avec les erreurs-types et les composantes de la variance pour les quatre méthodes. \* Valeur différentielle pour chaque ménage (moyenne = 0,1576633) \*\* Basé sur les résultats pour la méthode ELL

Variables	ELL(sans hétérosc.)	ELL(avec hétérosc.)	Pseudo-MPLSB	IIEFP	RGE
explicatives	Béta	Erreur-type	Béta	Erreur-type	Béta
ramsize	-0,11867	0,00181	-0,12034	0,00165	-0,11875
ramsizeqc	0,00093	0,00081	0,00036	0,00039	0,00038
type_mult	0,03876	0,01697	0,03703	0,01588	0,03699
per_kids	-0,20342	0,01476	-0,20818	0,01332	-0,20293
roof_light	-0,06314	0,01291	-0,05808	0,01056	-0,06263
roof_strong	-0,05882	0,01135	-0,05633	0,00962	-0,05448
wall_light	-0,05459	0,01182	-0,04979	0,00975	-0,05426
wall_salvaged	-0,10814	0,02505	-0,11377	0,02058	-0,10748
wall_strong	0,14248	0,01051	0,12964	0,00910	0,14274
fa_xs	-0,17052	0,00941	-0,16756	0,00782	-0,17144
fa_s	-0,08368	0,00861	-0,08242	0,00725	-0,08403
fa_l	0,09016	0,00908	0,08478	0,00792	0,09065
fa_xl	0,16959	0,01104	0,15404	0,00992	0,17034
fa_xxl	0,27072	0,01144	0,24485	0,01094	0,27172
all_eled	0,36190	0,01371	0,31369	0,01286	0,36270
all_hsed	0,42325	0,01250	0,43771	0,01083	0,42192
all_coed	1,21591	0,01371	1,29386	0,01379	1,21324
dom_help	0,60207	0,01629	0,61218	0,01886	0,60035
head_male	-0,05878	0,00988	-0,04581	0,00922	-0,05862
no_spouse	-0,09367	0,00987	-0,07376	0,00917	-0,09361
hea_rel_mus	0,28537	0,07654	0,25643	0,07375	0,28871
Per_eng	0,17273	0,06529	0,14561	0,06298	0,17782
Hou_coeipg	0,37463	0,04348	0,39784	0,04210	0,37934
Hou_own_ref	0,17716	0,10497	0,18342	0,10178	0,17189
Hou_own_tel	1,39287	0,13356	1,42109	0,12987	1,38551
Per_wor_pth	0,46957	0,15484	0,40302	0,14926	0,47517
Per_ind_52	-0,76245	0,21708	-0,78120	0,21073	-0,76326
const	9,54013	0,05525	9,54456	0,05290	9,53566
Estimation des composantes de la variance	Niveau de la grappe	Niveau de la grappe	Niveau de la grappe	Niveau de la grappe	Niveau de la grappe
0,18461	0,04741	NA*	0,04741	0,18820	0,05172
0,18461	0,04741	0,18461	0,04741	0,18185	0,05498
0,04741	la grappe**	Niveau de la grappe**	la grappe**	Niveau de la grappe	la grappe
0,04741	la grappe**	Niveau de la grappe**	la grappe**	Niveau de la grappe	la grappe



ensembles de données pouvant être fusionnés avec l'enquête de variables auxiliaires provenant de la FIES et de l'EPA.

Les tableaux 1, 2 et 3 donnent les estimations du paramètre ( $\beta$ ) et des erreurs-types correspondantes, ainsi que les estimations des composantes de la variance aux niveaux national, régional et provincial, respectivement. Le tableau 2 donne les résultats pour l'un des 16 modèles ajustés au niveau régional (il existait 16 régions aux Philippines en 2000). De même, le tableau 3 donne les résultats de l'un des 20 modèles provinciaux formulés pour les 20 provinces sélectionnées. Afin de normaliser les comparaisons, exactement le même ensemble de variables prédictives est utilisé pour toutes les méthodes d'ajustement du modèle. (Il existe cinq ensembles d'estimations du paramètre, bien que nous n'ayons examiné que quatre méthodes fondamentales, parce que la méthode ELL est utilisée avec et sans hétéroscédasticité.) Notons qu'en pratique, dans le cas de la méthode ELL, on subdivise souvent les données d'enquête et on ajuste des modèles distincts à chaque sous-échantillon, par exemple à chacune des strates définies selon les 16 régions des Philippines, voire même des modèles au niveau provincial. Cette approche peut produire des modèles surajustés et des erreurs-types présentant un biais par défaut pour les estimations sur petits domaines. Pour l'analyse présentée ici, nous n'avons ajusté qu'un seul modèle (c'est-à-dire le modèle au niveau national). En pratique, des modèles intermédiaires comportant certains effets régionaux, mais non tous, semblent être ceux qui donnent les meilleurs résultats. Voir par exemple Haslett et Jones (2005).

Afin d'évaluer les différences entre les estimations produites par les diverses méthodes, nous effectuons une comparaison informelle de la « signification » des diverses estimations de  $\beta$  en soustrayant de l'estimation produite par une méthode la moyenne des estimations obtenue par les autres méthodes, puis en divisant le résultat par l'erreur-type de la méthode en question. Au niveau national (tableau 1), les estimations des coefficients de régression produites par les diverses méthodes diffèrent de manière significative les unes des autres pour un certain nombre de variables indépendantes. La méthode RGE a tendance à produire, pour la majorité des variables, des estimations des coefficients de régression qui diffèrent de manière significative de celles obtenues par les autres méthodes. Comme nous l'avons souligné plus haut, l'estimateur RGE est l'estimateur par la régression pondérée par les poids de sondage pour un modèle avec structure de variance homoscedastique et observations non corrélées dans la population, de sorte que cet estimateur n'est pas dérivé sous le modèle spécifié par (7). Toutefois, il s'agit de l'estimateur le plus prudent, car il donne l'erreur-type la plus élevée pour toutes les caractéristiques au niveau du ménage. Par ailleurs, la méthode IBEF produit l'erreur-type estimée la plus grande

incidence sur les revenus et dépenses. Les ménages sélectionnés sont interviewés en deux opérations distinctes, couvrant chacune une période d'une demi-année, afin de tenir compte des variations saisonnières des revenus et des dépenses. Pour la FIES de 2000, les interviews ont été réalisées en juillet 2000 pour la période allant du 1<sup>er</sup> janvier au 30 juin, et en janvier 2001 pour la période allant du 1<sup>er</sup> juillet au 31 décembre. Le plan de sondage de la FIES s'appuyait sur l'échantillonnage aléatoire stratifié à plusieurs degrés. Les barangays, qui sont les unités primaires d'échantillonnage (LPE), sont répartis en une strate urbaine et une strate rurale dans chaque province et sélectionnés par échantillonnage systématique avec probabilité proportionnelle à la taille. Les grands barangays sont en outre subdivisés en secteurs de recensement et soumis à un échantillonnage supplémentaire avant l'étape finale durant laquelle les ménages sont échantillonnés systématiquement en se servant de la liste des ménages du Recensement de la population de 1995. La non-réponse à l'interview était de 3,4 % seulement, 39 615 ménages échantillonnés ayant pu être interviewés aux deux visites de l'enquête. La non-réponse partielle a été corrigée par imputation déterministe, c'est-à-dire qu'une entrée manquant pour une question particulière a été déduite d'après les réponses obtenues pour d'autres items du questionnaire.

Les variables auxiliaires utilisées dans le présent article sont celles incluses dans le modèle formulé par Haslett et Jones (2005), qui a été ajusté sans utiliser le logiciel POVMAP pour le projet de cartographie de la pauvreté par petit domaine aux Philippines. Les variables auxiliaires comprenaient les caractéristiques du ménage ainsi que les moyennes municipales (dans lesquelles les données sur les ménages utilisées ont la même valeur pour chaque ménage échantillonné dans une municipalité donnée, c'est-à-dire un petit domaine). Ces variables auxiliaires sont non seulement dérivées des données de la FIES, mais aussi de celles de l'Enquête sur la population active (EPA) de 2000 et du Recensement de la population et du logement (RPL) des Philippines. L'EPA est conçue pour recueillir des données sur les caractéristiques socioéconomiques de la population de plus de 15 ans. Le NSO la réalise trimestriellement par interview sur place en utilisant la semaine précédente comme période de référence. Comme elles faisaient partie de l'Integrated Survey of Households (NSCB 2000), les enquêtes de juillet 2000 et de janvier 2001 ont été réalisées auprès du même échantillon de ménages que la FIES de 2000. Donc, les deux ensembles de données peuvent être fusionnés pour former un ensemble plus riche de variables auxiliaires. Des variables supplémentaires ont également été tirées du Recensement de la population et du logement de 2000 sous forme de moyennes municipales. Les moyennes pour les variables de recensement du questionnaire abrégé ainsi que du questionnaire complet ont été calculées au niveau municipal pour créer de nouveaux



l'avons montré plus haut, les équations pour l'estimateur du paramètre  $\beta$  et sa matrice de covariance estimée correspondent ne font intervenir que la matrice de poids de sondage  $W$ . La matrice de covariance estimée donnée par (24) est souvent appelée estimateur sandwich.

## 6. Comparaison des méthodes d'ajustement de modèle

La méthode ELL est décrite comme une méthode d'estimation par moindres carrés généralisés (MCG) pondérée. Toutefois, comme nous l'avons souligné plus haut, les poids de sondage ne sont pas intégrés correctement dans le processus d'estimation, ce qui rend ininterprétables les éléments de certaines matrices intervenant dans l'estimation et rend asymétrique la matrice de covariance estimée. Dans la méthode ELL d'estimation des composantes de la variance, les poids ne sont pris en compte qu'au niveau de la grappe. Les deux moyens (calcul direct et calcul fondé sur un modèle d'hétéroscédasticité) utilisés par Elbers et coll. pour produire la composante de la variance au niveau du ménage  $n$  intègrent pas les poids de sondage. Dans le calcul direct, la composante de la variance au niveau du ménage est déterminée d'après le résidu de la régression pondérée par les poids de sondage (MCP) réalisée à l'étape préliminaire et l'estimation pondérée de la composante au niveau de la grappe. Le calcul fondé sur l'hétéroscédasticité repose sur la modélisation du carré des résidus de la régression par MCP. Alors que la méthode ELL suit une procédure d'estimation de type MCG, les méthodes pseudo-MPLSB et IEEF suivent la procédure des équations d'estimation généralisées (EEG) (Liang et Zeger 1986) avec utilisation d'une matrice de corrélation de travail échangeable, c'est-à-dire que tous les éléments hors diagonale de la matrice de corrélation à l'intérieur des grappes sont égaux, et dans le cas des méthodes pseudo-MPLSB et IEEF, sont égaux à  $\sigma_v^2/(\sigma_v^2 + \sigma_e^2)$ . La matrice de corrélation de travail échangeable ou équilibrée est l'une des matrices de corrélation de travail fréquentes présentées par Horton et Lipsitz (1999) dans un article où ils passent en revue divers projets pour l'ajustement de modèles de régression EEG. Les deux méthodes, pseudo-MPLSB et IEEF, intègrent toutes deux les poids de sondage dans l'estimation du paramètre  $\beta$  et de l'erreur-type correspondante, quoique la méthode pseudo-MPLSB recourt à la méthode de Henderson dans l'estimation des composantes de la variance. Alors que la méthode de Henderson produit des estimations non pondérées des composantes de la variance, la méthode IEEF intègre les poids de sondage itérativement en partant de l'estimation des composantes de la variance pour le calcul de l'erreur-type de l'estimation du paramètre de régression.

## 7. Application aux données réelles

Les publications portant sur l'application des méthodes pseudo-MPLSB et IEEF à des ensembles de données réelles sont très peu nombreuses. Celles qui existent traitent les grappes comme étant les petits domaines et s'appuient souvent sur l'ensemble de données présentées dans Bateese, Harter et Fuller (1988), qui contient des renseignements sur les hectares de maïs et de soja par segments pour les comtés du Centre-Nord de l'Iowa et repose sur l'hypothèse d'un échantillonnage aléatoire simple à l'intérieur des domaines ou des grappes. Fait exception l'article récent de Militino, Ugarte, Goicoa et Gonzalez-Audicana (2006), dans lequel le pseudo-MPLSB est appliqué pour estimer la superficie totale occupée par les oliviers à Navarra, en Espagne, où, comme dans Bateese et coll., les unités sont autopondérées. En général, pour l'estimation de la pauvreté, les méthodes pseudo-MPLSB et IEEF doivent être appliquées dans des situations plus complexes, puisque les grappes d'échantillonnage et les petits domaines ne sont pas identiques et que l'échantillon n'est pas autopondéré. Dans l'exemple de la section suivante, les grappes (barangays) diffèrent des petits domaines (municipalités), les grappes sont des sous-unités du petit domaine et le plan d'échantillonnage n'est pas autopondéré. La méthode de régression RGE est l'une des méthodes prises en considération, les procédures d'estimation fondées sur des données d'enquête pour le paramètre  $\beta$  et son erreur-type correspondante sont théoriquement valables étant donné les hypothèses sur lesquelles elles reposent, sauf dans le cas de la méthode ELL pour laquelle existent certaines incohérences dans l'estimation du paramètre  $\beta$  et de la covariance de  $\beta$ . Selon la discussion qui précède, pour toutes les méthodes prises en considération, les procédures d'estimation fondées sur des données d'enquête pour le paramètre  $\beta$  et son erreur-type correspondante sont théoriquement valables. À la présente section, nous comparons les quatre méthodes de régression étudiées (dont une contient deux variantes de la méthode ELL) en utilisant les données de la Family Income and Expenses Survey (FIES) de 2000 aux Philippines. La FIES est une enquête de portée nationale réalisée tous les trois ans par le National Statistics Office (NSO) des Philippines. L'enquête est conçue pour recueillir des données sur les revenus et les dépenses des familles, ainsi que des renseignements sur les facteurs ayant une

Les composantes de la variance sont estimées en utilisant la méthode 3 de Henderson (Henderson 1953) pour produire des estimations sans biais, même en présence d'éléments corrélés dans le modèle. Les estimateurs des composantes de la variance sont les suivants :

$$\hat{\sigma}_{\varepsilon^2}^{cH} = (n - k - p + 1)^{-1} \sum_{h=1}^a \sum_{n_h} \varepsilon_{ah}^2 \quad (19)$$

où  $\{\varepsilon_{ah}^2\}$  représente les résidus de la régression par les moindres carrés ordinaires (MCO) de  $(\bar{y}_{ah} - \bar{y}_a)$  sur  $\{x_{ah1}, \dots, x_{ahp} - \bar{x}_{a,p} - \bar{x}_{a,d}\}$  et  $(\bar{y}_a, \bar{x}_{a,1}, \dots, \bar{x}_{a,p})$  sont les moyennes d'échantillon dans le  $a^e$  groupe.

$$\hat{\sigma}_{\varepsilon^2}^{cH} = n_a^{-1} \left[ \sum_{h=1}^a \sum_{n_h} n_{ah}^2 - (n - p) \right] \hat{\sigma}_{\varepsilon^2}^{cH} \quad (20)$$

où  $n_a = n - \text{tr}[(X'X)^{-1} \sum_{a=1}^a n_a^2 \bar{x}_a \bar{x}_a']$  avec  $X = (x_1, \dots, x_k)$ , et les  $\{\bar{x}_{ah}\}$  sont les résidus de la régression par les MCO de  $y_{ah}$  sur  $\{x_{ah1}, \dots, x_{ahp}\}$ . Pour le modèle (7), l'indice inférieur  $a$  est remplacé par  $b$ .

Cependant, les estimateurs de Henderson susmentionnés ne tiennent pas compte des poids de sondage. Pour contourner ce problème, You et coll. (2003) ont proposé une technique d'estimation qui consiste à étendre la méthode du pseudo-MPLSB en intégrant les poids dans l'estimation des composantes de la variance, ce que nous décrivons à la section suivante.

## 5.2 La méthode itérative à équations d'estimation pondérées

L'estimateur proposé par You et coll. (2003) est semblable à l'estimateur pseudo-MPLSB, excepté qu'il intègre les poids de sondage dans le calcul des composantes de la variance, et produit l'estimation des paramètres  $\beta$  et des composantes de la variance en se fondant sur une approche itérative à équation d'estimation pondérées (IEBP). Les auteurs ont calculé les estimateurs de  $\sigma_{\varepsilon^2}^2$  et  $\sigma_v^2$  de la façon suivante :

$$\hat{\sigma}_{2(t)}^{c2(v)} = \frac{\sum_k \sum_{n_h} w_{ah}^2 [y_{ah} - \bar{y}_{ah} - (\bar{x}_{ah} - \bar{x}_{ah}^{aw})' \beta^{(t-1)}]^2}{\sum_h \left[ (1 - \delta_h^2) \sum_{n_h} w_{ah}^2 \right]}$$

$$\hat{\sigma}_{2(t)}^{c2(w)} = \beta^{(t)} \quad (21)$$

$$\hat{\sigma}_{2(t)}^{c2(w)} = \frac{1}{k} \sum_k \frac{k}{\sigma_{2(t-1)}^{c2(w)}} \sum_k \frac{k}{\sigma_{2(t-1)}^{c2(w)}} (\gamma_{aw} - 1)^2 + \frac{k}{\sigma_{2(t)}^{c2(w)}} \sum_k \frac{k}{\sigma_{2(t)}^{c2(w)}} \gamma_{2}^{aw} \quad (22)$$

Les estimations pondérées de  $\beta$ ,  $\sigma_{\varepsilon^2}^2$  et  $\sigma_v^2$  sont obtenues simultanément en suivant des étapes itératives de mise à jour,  $t$  représentant dans l'équation susmentionnée la  $t^e$  itération. Puisque les composantes de la variance  $\sigma_v^2$  et  $\sigma_{\varepsilon^2}^2$  sont inconnues, les estimations de départ pour les étapes d'itération

sont générées par la méthode de Henderson. De nouveau, comme pour le pseudo-MPLSB, dans la formule du modèle de régression ELL (7), l'indice inférieur  $a$  est remplacé par  $b$ .

Cette approche est semblable à la méthode des moindres carrés généralisés itérés pondérés par les probabilités de sélection (MCGPPS) proposée par Pfeffermann et coll. (1998) pour ajuster des modèles multivariés où le procédé d'estimation tient compte des probabilités de sélection inégales à chaque degré d'échantillonnage et comporte une itération entre le paramètre  $\beta$  et les composantes de la variance jusqu'à la convergence. Pfeffermann, Moura et Silva (2006) proposent aussi une approche fondée sur un modèle qui comprend le calcul du modèle hiérarchique pour des données d'échantillon données sous la forme d'une fonction du modèle de population et des probabilités de sélection, puis l'ajustement du modèle sur données d'échantillon selon une approche bayésienne en se servant de l'algorithme de Monte Carlo par chaîne de Markov.

## 5.3 Méthode de régression généralisée sur données d'enquête

Une autre approche pour produire l'estimateur du paramètre  $\beta$  et de sa variance est la méthode fondée sur le plan de sondage pour l'ajustement des modèles de régression (Lohr 1999). Cette technique est utilisée à l'heure actuelle dans les logiciels Stata, Sudaan et WesVar, par exemple. L'estimateur de  $\beta$  donné ci-après est l'estimateur par la régression pondérée par les poids de sondage pour un modèle avec structure de variance homoscedastique et observations non corrélées dans la population.

$$\hat{\beta}_S = (X'WX)^{-1} X'WY. \quad (23)$$

Cet estimateur n'est pas dérivé sous le modèle spécifié par (7), même sous l'hypothèse de variances homoscedastiques pour les erreurs au niveau du ménage. L'estimation linéarisée/robuste de variance pour  $\hat{\beta}_S$  est basée sur l'estimateur de variance sous le plan de sondage pour un total, donné par

$$V(\hat{\beta}_S) = D \left\{ \frac{m-1}{m} \sum_{n_h} w_{bh}^2 d_{bh} \right\} \left( \sum_{n_h} w_{bh}^2 d_{bh} \right)^{-1} D \quad (24)$$

où  $d_{bh} = \varepsilon_{bh}^2 x_{bh}^2$ ,  $\varepsilon_{bh}^2$  est le résidu de la régression par moindres carrés pondérés (MCP) ;  $x_{bh}$  est un vecteur de variables indépendantes ;  $w_{bh}$  est un poids de sondage ;  $D = (X'WX)^{-1}$  et  $W$  est une matrice diagonale de poids de sondage. La méthode de régression généralisée sur données d'enquête diffère des autres techniques en ce qui a trait au calcul des estimations et produit des estimations sans calcul des composantes de la variance,  $\sigma_v^2$  et  $\sigma_{\varepsilon^2}^2$ . Comme nous



que la première phase. La phase suivante comporte la prédiction au niveau du ménage en se basant sur les données de recensement complètes et l'agrégation au niveau du petit domaine.

Les méthodes d'ajustement aux données d'enquête (calcul de l'estimation de  $\beta$  et de sa matrice de variance-covariance correspondante) pour les trois méthodes de régression proposées comme alternative de la méthode ELL sont présentées aux sections qui suivent.

## 5. Autres méthodes d'ajustement

### 5.1 La méthode du meilleur prédicteur linéaire sans biais pseudo-empirique

You et Rao (2002) ont proposé un estimateur de la moyenne de petit domaine en établissant un estimateur de  $\beta$  basé sur le modèle au niveau de l'unité (4). L'établissement de l'estimateur de  $\beta$  débute par le calcul du meilleur prédicteur linéaire sans biais (MLSB) de  $v_a$  sachant les paramètres  $\beta$ ,  $\sigma_e^2$  et  $\sigma_v^2$  tirés du modèle au niveau du domaine agrégé (pondéré par les poids de sondage) :

$$\hat{Y}^{aw} = \mathbf{X}^{aw} \beta + v_a + \bar{e}^{aw} \quad (14)$$

qui procède comme il suit :

$$\hat{v}_a^{aw} (\beta, \sigma_e^2, \sigma_v^2) = \gamma^{aw} (\bar{Y}^{aw} - \mathbf{X}^{aw} \beta) \quad (15)$$

où  $\mathbf{X}^{aw} = \sum_{h=1}^a w_{ah}^{aw} \mathbf{X}^{ah}$ ,  $\bar{Y}^{aw} = \sum_{h=1}^a w_{ah}^{aw} \bar{Y}^{ah}$ ,  $\gamma^{aw} = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2 \delta_a^2)$ ,  $\delta_a^2 = \sum_{h=1}^a w_{ah}^{aw} w_{ah}^2$  et  $w_{ah}^{aw}$  sont les poids de sondage au niveau de l'unité ; vient ensuite la résolution de l'équation d'estimation pondérée par les poids de sondage pour trouver  $\beta$  :

$$\sum_{h=1}^a w_{ah}^{aw} \mathbf{X}^{ah} [\mathbf{Y}^{ah} - \mathbf{X}^{ah} \beta - \hat{v}_a^{aw} (\beta, \sigma_e^2, \sigma_v^2)] = 0 \quad (16)$$

d'après laquelle l'estimateur de  $\beta$  est obtenu sous la forme

$$\hat{\beta}_w = \left\{ \sum_{h=1}^a \sum_{n_a} \mathbf{X}^{ah} \mathbf{Z}^{ah} \right\}^{-1} \left\{ \sum_{h=1}^a \sum_{n_a} \mathbf{X}^{ah} \mathbf{Y}^{ah} \right\} \quad (17)$$

où  $\mathbf{Z}^{ah} = \mathbf{W}^{ah} (\mathbf{X}^{ah} - \gamma^{aw} \mathbf{X}^{aw})$ . La matrice de covariance correspondante s'écrit alors :

$$\Phi^w = \sigma_e^2 \left( \sum_{h=1}^a \sum_{n_a} \mathbf{X}^{ah} \mathbf{Z}^{ah} \right)^{-1} + \left( \sum_{h=1}^a \sum_{n_a} \mathbf{X}^{ah} \mathbf{Z}^{ah} \right)^{-1} \left( \sum_{h=1}^a \sum_{n_a} \mathbf{X}^{ah} \mathbf{Z}^{ah} \right)^{-1} \left( \sum_{h=1}^a \sum_{n_a} \mathbf{X}^{ah} \mathbf{Z}^{ah} \right)^{-1} \quad (18)$$

$$\ln \left[ \frac{e^{b_h} A^* - e^{b_h}}{e^{b_h}} \right] = \mathbf{Z}^{bh} \alpha + r_{bh} \quad (12)$$

d'estimer la forme plus simple

semblables des paramètres  $\alpha$ . Cette contrainte permet

$A^* = (1, 05) \max \{e^{b_h}\}$  produit en général des estimations

une borne minimale nulle et une borne maximale égale à

variables auxiliaires. Elbers et coll. soutiennent qu'imposer

semblance (Elbers et coll. 2003) et où  $\mathbf{Z}^{bh}$  représente les

sant la méthode classique du pseudo-maximum de vrais-

vement, estimées par le vecteur de paramètres  $\alpha$  en utili-

où A et B sont les bornes supérieure et inférieure, respecti-

où  $r_{bh}$  est un terme d'erreur et les autres variables sont les mêmes que celles définies précédemment. Dans la plupart des projets de cartographie de la pauvreté de la Banque mondiale, de légères modifications sont généralement apportées, par exemple l'ajout d'une constante  $\delta$  à  $e^{b_h}$  dans le modèle (11).

En utilisant le modèle (12), et en appliquant la méthode delta,  $\hat{\sigma}_{e^{bh}}^2$  est calculé sous la forme :

$$\hat{\sigma}_{e^{bh}}^2 = \left[ A^* C^{bh} (1 - C^{bh}) \right] \left[ \frac{1}{\hat{\sigma}_2^2} + \frac{1}{2} \hat{\sigma}_r^2 \right] \quad (13)$$

où  $C^{bh} = \exp \{ \mathbf{Z}^{bh} \alpha \}$ , et  $\hat{\sigma}_2^2$  est la variance estimée des résidus sous le modèle (12). Si la composante de la variance au niveau du ménage est fondée sur un modèle hétéroscédastique, alors  $\mathbf{V}^b = (\sigma_2^2 \mathbf{I}_{n_b} + \sigma_v^2 \mathbf{1}_{n_b} \mathbf{1}_{n_b}')$ . La modalisation de l'hétéroscédasticité est effectuée en supposant que la variation au niveau du ménage dépend de certaines

Come nous l'exposons plus en détail à l'annexe, la façon dont la matrice de pondération  $\mathbf{W}^b$  entre dans le calcul de l'équation (9) susmentionnée aboutit à une matrice de covariance estimée asymétrique. Une approche un peu meilleure basée sur le « pseudo-maximum de vraisemblance », qui est décrite par Pffeffermann, Skinner, Goldstein et Rasbash (1998), consiste à décomposer  $\mathbf{X}^b \mathbf{V}^{b-1} \mathbf{X}^b$  en sommes distinctes de carrés et de produits croisés, et à pondérer chacune de manière appropriée – si nous écrivons  $\mathbf{V}^{b-1} = c \mathbf{I}_{n_b} + d \mathbf{1}_{n_b} \mathbf{1}_{n_b}'$ , alors la pondération appropriée est  $c \mathbf{X}^b \mathbf{W}^b \mathbf{X}^b + d \mathbf{X}^b \mathbf{W}^b \mathbf{1}_{n_b} \mathbf{1}_{n_b}' \mathbf{W}^b \mathbf{X}^b$ .

Puisque la version ELL,  $\mathbf{W}^b \mathbf{V}^{b-1}$ , n'est en général pas symétrique, dans l'équation (9),  $\mathbf{D}$  ne l'est pas non plus. Donc, la matrice de covariance supposée de  $\hat{\beta}^{ELL}$ ,  $\mathbf{V}(\hat{\beta}^{ELL})$ , n'est pas symétrique non plus. Le logiciel POVMAP essaye de résoudre ce problème en prenant la moyenne de  $\mathbf{V}(\hat{\beta}^{ELL})$  et de sa transposée, ce qui force la matrice à être symétrique.

Mentionnons de nouveau que, sous la méthode ELL, l'ajustement de la régression aux données d'enquête et l'estimation des composantes de la variance ne constituent



note en bas de page 8) et dans le logiciel POVMAP développé pour la méthode ELL par Zhao (2006), sous la forme

$$\beta_{ELL} = \left( \sum_{m=1}^b \mathbf{X}_b^b \mathbf{V}_b^b \mathbf{V}_b^{-1} \mathbf{X}_b^b \right)^{-1} \left( \sum_{m=1}^b \mathbf{X}_b^b \mathbf{V}_b^b \mathbf{V}_b^{-1} \mathbf{y}_b \right) \quad (8)$$

et la matrice de variance-covariance correspondante sous la forme

$$\mathbf{V}(\beta_{ELL}) = \mathbf{D} \left[ \left( \sum_{m=1}^b \mathbf{X}_m^b \mathbf{V}_m^b \mathbf{V}_m^{-1} \mathbf{W}_m^b \mathbf{X}_m^b \right)^{-1} \right] \mathbf{D} \quad (9)$$

où  $\mathbf{V}^b = (\sigma_v^2 \mathbf{I}_{n_b} + \sigma_v^2 \mathbf{I}_{n_b} \mathbf{I}_{n_b}')$ ,  $(\sigma_v^2)$  est la variance au niveau de la grappe, tandis que  $(\sigma_e^2)$  est la variance au niveau du ménage,  $\mathbf{I}_{n_b}$  est la matrice identité,  $\mathbf{I}_{n_b} = (1 \dots 1)'$  est un vecteur constant,  $\mathbf{D} = (\sum_{m=1}^b \mathbf{X}_m^b \mathbf{V}_m^b \mathbf{V}_m^{-1} \mathbf{X}_m^b)^{-1}$ ,  $\mathbf{X}_m^b = (\mathbf{x}_{bm_1}^b, \dots, \mathbf{x}_{bm_{n_b}}^b)'$ ,  $\mathbf{y}_m^b = (\mathbf{y}_{bm_1}^b, \dots, \mathbf{y}_{bm_{n_b}}^b)'$ ,  $\mathbf{W}_m^b$  est une matrice diagonale des poids de sondage ;  $m$  est le nombre de grappes dans l'échantillon et  $n_b$  est le nombre de ménages dans chaque grappe échantillonnée. L'équation (8) repose sur l'hypothèse que  $\mathbf{V}_b^b$  est connue. En pratique, nous devons estimer  $\sigma_e^2$  et  $\sigma_v^2$  pour obtenir l'estimateur  $\hat{\mathbf{V}}_b^b$ . Nous notons que l'expression de la variance dans (9) est dérivée sous un modèle hypothétique vaguement spécifié pour l'échantillon (voir Elbers et coll. 2002). Sous la méthode ELL, l'ajustement du modèle de revenu/dépenses (7) comprend l'obtention de l'estimation initiale de  $\beta$  par la méthode des moindres carrés pondérés (MCP) et l'utilisation des résidus du modèle initial pour estimer la matrice de covariance  $\mathbf{V}_b^b$  nécessaire pour obtenir  $\beta_{ELL}$ . Les estimations des variances au niveau de la grappe  $(\sigma_v^2)$  et au niveau du ménage  $(\sigma_e^2)$  sont calculées par Elbers et coll. (2002) de la façon suivante :

$$\hat{\sigma}_v^2 = \max \left( \frac{\sum_b w_b (n_b - n_{..})^2}{\sum_b w_b (1 - w_b)} \tau_b^2, 0 \right) \quad (10)$$

où  $\tau_b^2 = \sum_b (e_{bh} - e_b)^2 / (n_b (n_b - 1))$  ;  $w_b = \sum_b w_{bh} / \sum_b \sum_b w_{bh}$  sont les poids de sondage transformés par grappe, dont la somme sur les grappes est égale à un, et  $w_{bh}$  représente les poids de sondage changés d'échelle, dont la somme est égale à la taille totale de l'échantillon. Ici,  $u_b = \sum_b u_{bh}$  et  $n_{..} = \sum_b \sum_b u_{bh}$  (qui est égal à zéro), où  $u_{bh}$  est défini comme dans l'équation (6).

Elbers et coll. (2002) ont proposé deux moyens de générer l'estimation de la composante de la variance au niveau du ménage, soit le calcul « direct » qui est désigné par  $(\hat{\sigma}_e^2)$  ou le calcul fondé sur un modèle hétéroscédastique  $(\hat{\sigma}_{e,bh}^2)$ . Le calcul direct s'appuie sur la différence entre l'erreur quadratique moyenne estimée d'après la régression MCP initiale et l'estimation calculée de  $\sigma_v^2$ , tandis que le calcul fondé sur un modèle hétéroscédastique s'appuie sur une fonction de lien de type logarithmique pour borner la variance comme il suit :

$$\sigma_{e,bh}^2(z_{bh}^e, \alpha, A, B) = \left[ \frac{1 + \exp(z_{bh}^e \alpha)}{\exp(z_{bh}^e \alpha) + B} \right] \quad (11)$$

de ménages dans la  $b^e$  grappe dans population.  $\mathbf{x}_{bh}^e$  est un ensemble de variables auxiliaires disponibles dans l'enquête ainsi que dans le recensement, qui doivent généralement être contemporaux :  $u_{bh}^e$  est le terme d'erreur aléatoire représentant la part de  $\mathbf{x}_{bh}^e$  qui ne peut pas être expliquée par  $\mathbf{x}_{bh}^e$ . Les données sur le revenu et les dépenses possédant presque invariablement une distribution asymétrique, une transformation (habituellement logarithmique) est appliquée pour les rendre plus symétriques.

Les ménages pour lesquels des données sur le revenu ou les dépenses par tête sont recueillies sont rarement indépendants, et forment des grappes naturelles, souvent définies administrativement. Les ménages qui sont proches les uns des autres ou appartiennent à la même grappe ont tendance à être similaires à de nombreux égards. Dans les données d'enquête, les grappes sont habituellement aussi les unités primaires d'échantillonnage (UPÉ) du plan de sondage. Afin de tenir compte de la mise en grappe des ménages, on suppose habituellement que le terme d'erreur aléatoire  $u_{bh}^e$  du modèle de régression a la spécification suivante :

$$u_{bh}^e = v_b + e_{bh} \quad (6)$$

où  $v$  et  $e$  sont indépendants l'un de l'autre et non corrélés à  $\mathbf{x}_{bh}^e$ ,  $v_b$  est le terme d'erreur appartenant en commun au  $b^e$  groupe ou grappe (par exemple barangay pour les Philippines) et  $e_{bh}$  est l'erreur au niveau du ménage à l'intérieur de la grappe. L'importance de ces termes est mesurée par leur variance ou les composantes de leur variance,  $\sigma_v^2$  et  $\sigma_e^2$ , respectivement. Diverses méthodes existent pour estimer ces variances. Nous abordons ce sujet important aux sections qui suivent.

Le modèle (5) peut s'écrire

$$\mathbf{y}_{bh}^e = \mathbf{x}_{bh}^e \beta^b + v_b + e_{bh} \quad (7)$$

dont la forme est similaire à celle du modèle au niveau de l'unité ou du modèle de régression à erreurs emboîtées mentionné à la section précédente. Cependant, si la forme du modèle est similaire, le groupe auquel il est fait référence est différent, par exemple  $\mathbf{x}_{bh}^e$  renvoie au  $h^e$  ménage dans le  $a^e$  petit domaine, tandis que  $\mathbf{x}_{bh}^e$  renvoie au  $h^e$  ménage dans la  $b^e$  grappe. Les grappes, fondées sur le plan de sondage, sont habituellement nettement plus petites que les domaines pour lesquels des estimations sur petits domaines sont recherchées et, généralement, elles ne sont pas toutes échantillonnées (contrairement aux petits domaines qui le sont presque tous). Par exemple, aux Philippines, des estimations sont demandées au niveau municipal, qui comprend les barangays ou grappes.

#### 4. La méthode ELL

Pour la méthode ELL, l'estimation du paramètre de régression  $\beta$  est donnée dans Elbers et coll. (2002, page 11,

où  $\theta_a = g(\tilde{Y}_a)$  et les erreurs dues à l'échantillonnage  $e_a$  sont indépendantes et suivent une loi  $N(0, V_a)$  de variance connue  $V_a$ . En combinant les équations (1) et (2), on obtient le modèle mixte linéaire au niveau du domaine :

$$\theta_a = \mathbf{x}_a' \boldsymbol{\beta} + c_a v_a + e_a. \quad (3)$$

Souignons que (3) fait intervenir à la fois des variables

aléatoires fondées sur le plan de sondage  $e_a$  et des variables aléatoires fondées sur le modèle  $v_a$  (Rao 1999), où les variables fondées sur le plan de sondage dépendent du mécanisme de sélection d'échantillons et celles fondées sur le modèle, de la structure de superpopulation dans laquelle le modèle est intégré.

Les modèles au niveau du domaine possèdent diverses extensions qui permettent, par exemple, de traiter des erreurs dues à l'échantillonnage corrélées, la dépendance spatiale des effets aléatoires de petit domaine, les séries chronologiques et les données transversales (voir Rao 2003, 1999, ainsi que Ghosh et Rao 1994).

Le modèle au niveau de l'unité repose sur l'hypothèse que la variable d'intérêt  $Y_{ah}$  pour la  $h^e$  unité dans le  $a^e$  petit domaine est reliée aux données auxiliaires propres à l'élément  $\mathbf{x}_{ah} = (x_{ah1}, \dots, x_{ahp})'$  par la voie d'un modèle de régression à erreurs emboîtées :

$$Y_{ah} = \mathbf{x}_{ah}' \boldsymbol{\beta} + v_a + e_{ah} \quad (4)$$

où  $a = 1, \dots, k$ ,  $h = 1, \dots, N_a$ ,  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p-1})'$  est le vecteur de dimensions  $p \times 1$  des paramètres de régression et  $N_a$  est le nombre d'unités ou de ménages de la population dans le  $a^e$  petit domaine. Il est également supposé que les effets aléatoires,  $v_a$ , sont  $\text{iid } N(0, \sigma_v^2)$  et sont indépendants des erreurs au niveau de l'unité,  $e_{ah}$ , qui sont supposées être  $\text{iid } N(0, \sigma_e^2)$ . Des extensions permettant que les erreurs soient hétéroscédastiques, avec constante(s) d'échelle connue(s) sont également possibles.

La méthode ELL s'appuie sur un modèle au niveau de l'unité, où les unités sont les ménages dans le cas des données sur les revenus ou les dépenses, et où la variation est modélisée au niveau de l'unité primaire d'échantillon-nage, c'est-à-dire au niveau de la grappe et au niveau du ménage. Notons que ELL n'intègre pas la variation du modèle au niveau du petit domaine, et ne le font que pour la grappe dans le petit domaine, et pour le ménage dans la grappe. Cette forme du modèle de base est celle utilisée pour les comparaisons dans le présent article, puisque la méthode ELL est la méthode standard d'estimation sur petits domaines de la pauvreté dans les pays du tiers monde. Dans les ensembles de données réelles que nous avons étudiés, cette variation supplémentaire au niveau du petit domaine était très faible. Toutefois, malgré ces preuves empiriques, d'importantes questions persistent quant à la meilleure façon d'estimer la composante de la variance au

### 3. Modèle de revenu/consommation

La réponse à la question de savoir quel est le « meilleur ajustement du modèle de régression » aux données d'enquête analysées dans le présent article (comme celles à d'autres questions liées à la méthode ELL) est particulière-ment importante, parce que des milliards de dollars destinés au financement de l'aide sont (ou pourraient être) répartis en se fondant sur les modèles de régression utilisés dans le cadre de l'estimation sur petits domaines de la pauvreté.

La modélisation du revenu ou des dépenses par tête dans les ménages au lieu de mesures de la pauvreté et l'écart de la dite(s) (telles que la prévalence de la pauvreté et l'écart de la pauvreté) est l'une des caractéristiques distinctives de la méthode ELL. Comme nous l'avons mentionné à la section précédente, cette méthode comprend l'ajustement du modèle de revenu ou de dépenses aux données d'enquête et l'application de ce modèle à des données de recensement avant de produire les estimations sur petits domaines des mesures de la pauvreté. Le modèle de revenu/dépenses est de la forme :

$$Y_{bh} = \mathbf{x}_{bh}' \boldsymbol{\beta} + u_{bh} \quad (5)$$

où  $b = 1, \dots, M$ ,  $h = 1, \dots, N_b$ ,  $Y_{bh}$  représente le revenu ou la  $h^e$  unité ou ménage dans la  $b^e$  grappe,  $M$  est le nombre total de grappes dans la population et  $N_b$  est le nombre total

niveau du petit domaine en présence de variations au niveau de la grappe, en cas de pondération par les poids de sondage, surtout quand de nombreux petits domaines ne contiennent qu'une seule grappe échantillonnée.



dans de telles situations que l'estimation sur petits domaines entre en jeu.

La méthode d'estimation sur petits domaines utilisée le plus fréquemment pour mesurer la pauvreté dans les pays du tiers monde, qui a été proposée par Elbers, Lanjouw et Lanjouw (EL) (2002, 2003), permet de produire des estimations plus précises pour les régions géographiques plus petites en combinant des données d'enquête avec des renseignements tirés d'un recensement récent. La méthode ELL comporte deux phases, à savoir l'ajustement d'un modèle (ou de modèles) de régression à des données d'enquête complexes, puis l'utilisation de ce modèle pour prédire le revenu et les dépenses par tête au niveau du ménage (valeurs qui sont transformées et agrégées pour estimer les statistiques sur la pauvreté au niveau du petit domaine).

Dans le présent article, nous nous penchons spécifiquement sur les divers algorithmes utilisés pour ajuster les modèles de régression à la première phase et pour estimer les erreurs-types des paramètres de régression et les composantes de la variance d'après des données d'enquête. Nous mettons l'accent sur les conséquences des décisions concernant la modélisation de régressions sur données d'enquête plutôt que sur le système entier et assez complet qu'ELL utilisent pour produire des estimations sur petits domaines.

L'exigence préliminaire, lorsque l'on applique la méthode ELL à des mesures économiques, est d'élaborer un modèle exact du revenu ou des dépenses par tête dans les ménages, bien que ce modèle soit souvent utilisé pour générer des fonctions non linéaires du revenu ou des dépenses (par exemple prévalence de la pauvreté – pourcentage de ménages sous le seuil de pauvreté, ou écart de pauvreté – somme des écarts relatifs entre le seuil de pauvreté et le revenu ou les dépenses des ménages ou des individus qui se trouvent sous le seuil de pauvreté). Le modèle de régression sur données d'enquête élaboré pour le revenu ou les dépenses joue un rôle essentiel dans la production de statistiques exactes sur la pauvreté, mais, comme nous le montrons plus loin, le modèle de régression proprement dit n'est pas toujours l'élément le plus important et d'autres questions, telles que l'estimation des composantes de la variance, méritent qu'on s'y attarde.

D'autres méthodes de régression sur données d'enquête pour l'estimation sur petits domaines – la méthode (pseudo-MPLSB) (You et Rao 2002), la méthode itérative à du meilleur prédicteur linéaire sans biais pseudo-empirique (pseud-MPLSB) (You et Rao 2002), la méthode itérative à équations d'estimations pondérées (IEEP) (You et coll. 2003) et la méthode de régression généralisée sur données d'enquête (RGE) (Skinner, Holt et Smith 1989) sont examinées en tant que techniques d'ajustement de modèles à des données d'enquête et comparées à deux variantes de la méthode ELL d'ajustement de modèles de régression à des

données d'enquête. Notre étude est fondée sur des données réelles tirées de la Family Income and Expenses Survey (FIES) des Philippines, plutôt que sur des données simulées. Le plan de l'article est le suivant. À la section 2, nous donnons des renseignements généraux sur les modèles d'estimation sur petits domaines; à la section 3, nous présentons le modèle pour le revenu (ou les dépenses) décrit par Elbers, Lanjouw et Lanjouw; à la section 4, nous présentons un sommaire de la méthode ELL; à la section 5, nous décrivons en détail les diverses méthodes d'ajustement, dont la méthode du meilleur prédicteur linéaire sans biais pseudo-empirique (5.1), la méthode IEEP (5.2) et la méthode de régression généralisée sur données d'enquête (5.3). À la section 6, nous discutons des différences entre les méthodes, tandis qu'à la section 7, nous présentons l'application des méthodes aux données de la FIES de 2000 des Philippines. Enfin, à la section 8, nous présentons nos conclusions et nos recommandations.

## 2. Modèles pour petits domaines

Ghosh et Rao (1994) classent les modèles pour petits domaines en deux grandes catégories, c'est-à-dire les modèles au niveau du domaine et les modèles au niveau de l'unité. Les modèles au niveau du domaine correspondent à l'ensemble de modèles qui peuvent être pris en considération quand on ne dispose que de variables auxiliaires propres au domaine. Les modèles au niveau de l'unité, par ailleurs, englobent ceux qui peuvent être considérés lorsqu'il existe des variables auxiliaires propres à l'unité et que des valeurs de la variable étudiée au niveau de l'unité peuvent être utilisées. Tous ces modèles sont des cas particuliers d'un modèle linéaire généralisé ou d'un modèle mixte linéaire généralisé et contiennent habituellement des effets fixes ainsi que des effets aléatoires.

Pour les modèles au niveau du domaine, on suppose que la moyenne de population ( $\bar{Y}_a$ ) du  $a^e$  petit domaine ou une fonction appropriée  $\theta_a = g(\bar{Y}_a)$  est reliée aux variables auxiliaires propres au domaine  $\mathbf{x}_a = (x_{a1}, \dots, x_{ap})'$  au moyen d'un modèle linéaire

$$\theta_a = \mathbf{x}_a' \boldsymbol{\beta} + c_a v_a \quad (1)$$

où  $a = 1, \dots, k$ ,  $v_a \sim iid(0, \sigma_v^2)$ ,  $\boldsymbol{\beta}$  est un vecteur de paramètres de régression,  $c_a$  est une constante positive connue ou estimée pour tenir compte de l'hétéroscédasticité,  $k$  est le nombre total de petits domaines d'intérêt et  $p$  est le nombre de variables auxiliaires. On suppose qu'un estimateur direct fondé sur le plan de sondage,  $\bar{Y}_a$ , de la moyenne de population  $\bar{Y}_a$  est disponible quand la taille de l'échantillon de domaine  $n_a \geq 1$ , et que

$$\theta_a = \theta_a + e_a \quad (2)$$



# Comparaison de méthodes de régression sur données d'enquête dans le contexte de l'estimation de la pauvreté sur des petits domaines

Stephen J. Haslett, Marissa C. Isidro et Geoffrey Jones<sup>1</sup>

## Résumé

L'une des clés de la réduction ou de l'éradication de la pauvreté dans le tiers monde est l'obtention d'information fiable sur les pauvres et sur leur emplacements, afin que les interventions et l'aide soient dirigées vers les personnes les plus nécessiteuses. L'estimation sur petits domaines est une méthode statistique utilisée pour surveiller la pauvreté et décider de la répartition de l'aide de façon à réaliser les Objectifs du millénaire pour le développement. Eibers, Lanjouw et Lanjouw (EL) (2003) ont proposé, pour produire des mesures de la pauvreté fondées sur le revenu ou sur les dépenses, une méthode d'estimation sur petits domaines qui est mise en œuvre par la Banque mondiale dans ses projets de cartographie de la pauvreté grâce à la participation des organismes statistiques centraux de nombreux pays du tiers monde, dont le Cambodge, le Laos, les Philippines, la Thaïlande et le Vietnam, et qui est intégrée dans le logiciel PovoMap de la Banque mondiale. Dans le présent article, nous présentons la méthode ELL, qui consiste à modéliser d'abord les données d'enquête, puis à appliquer le modèle obtenu à des données de recensement, en nous penchant surtout sur la première phase, c'est-à-dire l'ajustement des modèles de régression, ainsi que sur les erreurs-types estimées à la deuxième phase. Nous présentons d'autres méthodes d'ajustement de modèles de régression, telles que la régression généralisée sur données d'enquête (RGE) (décrite dans Lohr (1999), chapitre 11) et celles utilisées dans les méthodes existantes d'estimations sur petits domaines, à savoir la méthode du meilleur prédicteur linéaire sans biais pseudo-empirique (pseudo-MPLSB) (You et Rao 2002) et la méthode itérative à équations d'estimation pondérées (IEEP) (You, Rao et Kovachev 2003), et nous les comparons à la stratégie de modélisation de ELL. La différence la plus importante entre la méthode ELL et les autres techniques tient au fondement théorique de la méthode d'ajustement du modèle proposée par ELL. Nous nous servons d'un exemple fondé sur la Family Income and Expenses Survey des Philippines pour illustrer les différences entre les estimations des paramètres et leurs erreurs-types correspondantes, ainsi qu'entre les composantes de la variance générées par les diverses méthodes et nous étendons la discussion à l'effet de ces différences sur l'exactitude des estimations sur petits domaines finales. Nous mettons l'accent sur la nécessité de produire de bonnes estimations des composantes de la variance, ainsi que des coefficients de régression et de leurs erreurs-types aux fins de l'estimation sur petits domaines de la pauvreté.

Mots clés : Modèles pour petits domaines ; modèle de régression à erreurs emboîtées ; cartographie de la pauvreté.

## 1. Introduction

La pauvreté est un problème multidimensionnel très complexe pour lequel il n'existe pas de définition unique ni de méthode de mesure unique. Dans le présent article, nous adoptions le sens donné au terme pauvreté par la plupart des économistes. Les ménages considérés comme étant en état de pauvreté sont ceux dont le revenu est inférieur à un certain seuil de revenu appelé seuil de pauvreté. Chambers (2006) donne à cette approche le nom de pauvreté fondée sur le revenu et cette définition est celle adoptée par la Banque mondiale dans la mise en œuvre de ses projets de cartographie de la pauvreté par petit domaine exécutés en collaboration avec les organismes statistiques nationaux et utilisés, par exemple, pour surveiller les progrès réalisés en regard des Objectifs du millénaire pour le développement (site Web de l'ONU). Parfois, des mesures de la pauvreté fondées sur les dépenses sont utilisées à la place de celles fondées sur le revenu pour évaluer la pauvreté économique. Dans les contextes de santé publique, diverses mesures, telles que le poids et la taille normalisés selon l'âge, ainsi

que le poids en fonction de la taille chez les enfants (insuffisance pondérale, retard de croissance et émaciation, respectivement) sont utilisés, par exemple, au Bangladesh (Haslett et Jones 2004) et au Népal (Haslett et Jones 2006). Les enquêtes réalisées dans la plupart des pays du tiers monde permettent habituellement d'obtenir un niveau acceptable de précision pour la publication de statistiques sur la pauvreté aux premier et deuxième niveaux administratifs ou au niveau de la région géographique (par exemple pour les Philippines, national et régional, respectivement). Cependant, pour que les responsables de l'élaboration des politiques puissent diriger correctement l'aide et les interventions vers les communautés et les ménages qui en ont le plus besoin, ils doivent disposer de statistiques sur la pauvreté produites à un niveau plus fin de détail. Toutefois, les statistiques sur la pauvreté calculées au moyen de données d'enquête pour de plus petites régions géographiques ou un plus faible niveau administratif sont habituellement moins fiables (possèdent des erreurs-types plus élevées) à cause des tailles d'échantillon plus petites et c'est

1. Stephen J. Haslett, Marissa C. Isidro et Geoffrey Jones, Institute of Fundamental Sciences: Statistics, College of Sciences, Massey University, Private Bag 11-222, Palmerston North, Nouvelle-Zélande. Courriel : S.J.Haslett@massey.ac.nz.

- Kim : Estimation par calage en utilisant l'inclinaison exponentielle dans les enquêtes par sondage
- Kott, P.S. (2003). A practical use for instrumental-variable calibration. *Journal of Official Statistics*, 19, 265-272.
- Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 149-160.
- Kitamura, Y., et Stutzert, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65, 861-874.
- Park, M., et Fuller, W.A. (2009). The mixed model for survey regression estimation. *Journal of Statistical Planning and Inference*, 139, 1320-1331.
- Rao, J.N.K., et Singh, A. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 57-64.
- Wolter, K.M. (2007). *Introduction to Variance Estimation*. 2<sup>ème</sup> Ed. New York : Springer-Verlag.
- Wu, C., et Rao, J.N.K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*, 34, 359-375.
- Tillé, Y. (1998). Estimation in surveys using conditional probabilities: Simple random sampling. *Revue Internationale de Statistique*, 66, 303-322.
- Sämdal, C.-E., Swenson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York : Springer.
- Sämdal, C.-E. (2007). La méthode de calage dans la théorie et la pratique des enquêtes. *Techniques d'enquête*, 33, 113-135.
- Rust, K.F., et Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

En utilisant  $w_i(\eta_{0,N}) = d_i$  et en écrivant  $\mathbf{s}_i(\eta_{0,N}) = \mathbf{s}_{i0}$ , nous avons, en vertu de (A2),

$$\frac{\partial}{\partial \eta} \frac{1}{N} \sum_{i \in A} w_i(\eta_{0,N}) y_i = -\frac{1}{N} \sum_{i \in A} d_i s_{i0} y_i$$

$$= -\hat{\Sigma}_{sy} + O_p(n^{-1/2}). \quad (A5)$$

En utilisant (A5) et (A3) dans (A4), nous obtenons le résultat (9).

## B. Preuve du théorème 2

Écrivons

$$\hat{\theta}(\lambda_1) = \frac{\sum_{i \in F} d_i m_i(\lambda_1) y_i}{\sum_{i \in I} d_i m_i(\lambda_1) y_i},$$

où  $m_i(\lambda_1) = \exp(\lambda_1' x_i)$ . Notons que  $X_{\text{ET}(i)} = N \hat{\theta}(\lambda_{1(i)})$  et que  $\lambda_{1(i)}$  est défini dans (19). Par un développement en série de Taylor de  $\hat{\theta}(\lambda_{1(i)}) = N^{-1} X_{\text{ET}(i)}$  autour de  $\lambda_1 = \mathbf{0}$  et en vertu de la continuité des dérivées partielles de  $\hat{\theta}(\lambda_1)$ , nous avons

$$\hat{\theta}(\lambda_{1(i)}) = \hat{\theta}(\mathbf{0}) + \hat{\theta}(\mathbf{0})'(\lambda_{1(i)} - \mathbf{0}) + o_p(|\lambda_{1(i)} - \mathbf{0}|). \quad (B1)$$

où  $\hat{\theta}(\lambda) = \partial \hat{\theta}(\lambda) / \partial \lambda$ . Comme la convergence de  $\lambda_{1(i)}$  est d'ordre quadratique et que l'estimateur en une étape satisfait  $\lambda_{1(i)} = O_p(n^{-1/2})$ , l'équation (22) peut s'écrire

$$\lambda_{1(i)} = \left\{ \hat{N}_i^{-1} \sum_{i \in A} d_i (x_i - \bar{X}_d) \right\}^{(32)} (\hat{N}^{-1} X - \bar{X}_d)$$

$$+ o_p(n^{-1/2}). \quad (B2)$$

Notons que

$$\hat{\theta}(\lambda_1) = \left\{ \sum_{i \in A} d_i m_i(\lambda_1) \right\}^{-1} \sum_{i \in A} d_i m_i(\lambda_1) y_i - \hat{\theta}(\lambda_1);$$

où  $m_i(\lambda_1) = \partial m_i(\lambda_1) / \partial \lambda_1$ . En utilisant  $m_i(\mathbf{0}) = 1$  et  $m_i'(\mathbf{0}) = x_i$ , nous avons  $\hat{\theta}(\mathbf{0}) = \bar{Y}_d / N_d$  et

$$\hat{\theta}(\mathbf{0}) = \hat{N}_d^{-1} \sum_{i \in A} d_i (x_i - \bar{X}_d) y_i. \quad (B3)$$

Par conséquent, en insérant (B2) et (B3) dans (B1), nous avons

$$\frac{\hat{Y}_d}{\hat{N}_d} = \hat{\theta}(\lambda_{1(i)}) +$$

$$+ \left( \frac{\bar{N}_d}{\bar{X}_d} - \bar{X}_d \right)' \left\{ \sum_{i \in A} d_i (x_i - \bar{X}_d) \right\}^{\otimes 2} \sum_{i \in A} d_i (x_i - \bar{X}_d) y_i$$

ce qui prouve (23).

## Bibliographie

- Beaumont, J.-F., et Bocci, C. (2008). Another look at ridge calibration. *Metron*, LXVI, 5-20.
- Breidt, F.J., Claeskens, G. et Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92, 831-846.
- Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.
- Chen, J., et Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.
- Chen, J., et Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.
- Chen, J., Varyath, A.M. et Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, 17, 426-443.
- Deville, J.-C., et Sâmdal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Deville, J.-C., Sâmdal, C.-E. et Sautory, O. (1993). Generalized raking procedure in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9, 139-172.
- Estévez, V.M., et Sâmdal, C.-E. (2000). A functional approach to calibration. *Journal of Official Statistics*, 16, 379-399.
- Folsom, R.E. (1991). Exponential and logistic weight adjustment for sampling and nonresponse error reduction. Dans *Proceedings of the Section on Social Statistics*, American Statistical Association, 197-202.
- Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.
- Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, New Jersey : John Wiley & Sons, Inc.
- Givens, G.H., et Hoeting, J.A. (2005). *Computational Statistics*. Hoboken, New Jersey : John Wiley & Sons, Inc.
- Hennri, M., Yoshida, R. et Eguchi, S. (2007). Importance sampling via the estimated sampler. *Biometrika*, 94, 985-991.
- Imbens, G.W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20, 493-506.
- Isaki, C., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- Kim, J.K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, 19, 145-157.
- Kim, J.K., et Park, M. (2010). Calibration estimation in survey sampling. *Revue Internationale de Statistique*, Sous presse.



où  $Y^{(0)} = \sum_{i \in \mathcal{A}} w_i^{(0)} Y_i$  et  $\sum_{i \in \mathcal{A}} w_i^{(0)} \{U(\mathbf{x}_i) - \bar{U}^{(0)}\} Y_i$ . Contrairement à l'estimateur par la régression, les poids de la méthode proposée sont toujours non négatifs. En outre, en utilisant la technique de la variable instrumentale décrite à la section 3, les poids possèdent une borne supérieure. Le choix approprié de la variable instrumentale accroît également l'efficacité de l'estimateur par calage résultant.

La méthode de calage par inclinaison exponentielle est asymptotiquement équivalente à la méthode de calage par la vraisemblance empirique, mais elle est plus intéressante du point de vue des calculs, en ce sens que les dérivées partielles ne sont pas requises dans le calcul itératif. Comme les calculs sont simples, la variance de l'estimateur proposé peut être estimée facilement en utilisant une méthode de rééchantillonnage, comme celle décrite à la section 4. Une étude plus approfondie de cette approche, y compris l'estimation par intervalle, pourrait être le sujet de futurs travaux de recherche.

## Remerciements

L'auteur remercie Minsun Kim de son soutien pour les calculs, ainsi que les deux examinateurs anonymes et le rédacteur associé de leurs commentaires très utiles qui lui ont permis d'améliorer considérablement la qualité de l'article. La présente étude a été financée en partie par l'entente de coopération NCRS 68-3A75-4-122 conclue entre le Natural Resources Conservation Service du US Department of Agriculture et la Iowa State University. Toutes les opinions, constatations et conclusions ou recommandations exprimées dans le présent article sont celles de l'auteur et ne reflètent pas forcément celles du USDA Natural Resources Conservation Service.

## Annexe

### A. Hypothèses et preuve du théorème 1

Pour commencer, supposons qu'existent les conditions de régularité suivantes :

[A-1] La densité  $f(\mathbf{x}; \boldsymbol{\eta})$  est deux fois dérivable par rapport à  $\boldsymbol{\eta}$  pour chaque  $\mathbf{x}$  et satisfait

$$\left| \frac{\partial^2 f(\mathbf{x}; \boldsymbol{\eta})}{\partial \eta_i \partial \eta_j} \right| \leq K(\mathbf{x})$$

pour la fonction  $K(\mathbf{x})$ , telle que  $E\{K(\mathbf{x})\} < \infty$ , dans un voisinage de  $\boldsymbol{\eta}_{0,N}$ .

[A-2] L'estimateur du pseudo-maximum de vraisemblance  $\hat{\boldsymbol{\eta}}$  satisfait  $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}) = O_p(1)$ .

[A-3] La matrice  $E\{\mathbf{s}(\boldsymbol{\eta}_{0,N})\}^{\otimes 2}$  existe et est non singulière, où  $\mathbf{s}(\boldsymbol{\eta}_{0,N}) = \partial \ln f(\mathbf{x}_i; \boldsymbol{\eta}) / \partial \boldsymbol{\eta} |_{\boldsymbol{\eta}=\boldsymbol{\eta}_{0,N}}$ .

Pour prouver le théorème 1, écrivons

$$g_i(\boldsymbol{\eta}) = \frac{f(\mathbf{x}_i; \boldsymbol{\eta}_{0,N})}{f(\mathbf{x}_i; \boldsymbol{\eta})},$$

et  $w_i(\boldsymbol{\eta}) = d_i g_i'(\boldsymbol{\eta})$ . Le poids d'échantillonnage préférentiel estimé donné par (8) peut s'écrire  $w_i = w_i(\hat{\boldsymbol{\eta}})$ . Un développement en série de Taylor de  $N^{-1} \sum_{i \in \mathcal{A}} d_i s_i(\hat{\boldsymbol{\eta}}) = 0$  autour de  $\boldsymbol{\eta}_{0,N}$  mène à

$$0 = \frac{1}{N} \sum_{i \in \mathcal{A}} d_i s_i(\boldsymbol{\eta}_{0,N}) + \left\{ \frac{\partial}{\partial \boldsymbol{\eta}} \frac{1}{N} \sum_{i \in \mathcal{A}} d_i s_i(\boldsymbol{\eta}_{0,N}) \right\} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}) + o_p(|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}|).$$

Notons que le premier terme du deuxième membre de

$$\frac{1}{N} \frac{\partial}{\partial \boldsymbol{\eta}} \sum_{i \in \mathcal{A}} d_i s_i(\boldsymbol{\eta}) = \frac{1}{N} \sum_{i \in \mathcal{A}} d_i \frac{\partial^2 f(\mathbf{x}_i; \boldsymbol{\eta}) / \partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'}{f(\mathbf{x}_i; \boldsymbol{\eta})}$$

$$- \frac{1}{N} \sum_{i \in \mathcal{A}} d_i \left\{ \frac{\partial f(\mathbf{x}_i; \boldsymbol{\eta}) / \partial \boldsymbol{\eta}}{f(\mathbf{x}_i; \boldsymbol{\eta})} \right\}^{\otimes 2}. \quad (\text{A1})$$

converge vers  $\int \{\partial^2 f(\mathbf{x}; \boldsymbol{\eta}) / \partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'\} d\mathbf{x}$  qui est égal à zéro en vertu du théorème de convergence dominée avec [A1]. Le deuxième terme converge vers  $E\{\mathbf{s}(\boldsymbol{\eta}_{0,N})\}^{\otimes 2}$ . Donc, en vertu de [A-2],

$$\bar{\mathbf{S}}_{0d} \equiv \frac{1}{N} \sum_{i \in \mathcal{A}} d_i s_i(\boldsymbol{\eta}_{0,N}) = O_p(n^{-1/2}) \quad (\text{A2})$$

et

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N} = \bar{\mathbf{S}}_{0d}^{-1} \bar{\mathbf{S}}_{0d}^{ss} + o_p(n^{-1/2}). \quad (\text{A3})$$

Or, un développement en série de Taylor de  $N^{-1} \bar{Y}_w = N^{-1} \sum_{i \in \mathcal{A}} w_i(\hat{\boldsymbol{\eta}}) Y_i$  autour de  $\boldsymbol{\eta} = \boldsymbol{\eta}_{0,N}$  mène à

$$\frac{\bar{Y}_w}{\bar{Y}} = \frac{N}{Y}$$

$$+ \left\{ \frac{\partial}{\partial \boldsymbol{\eta}} \frac{1}{N} \sum_{i \in \mathcal{A}} w_i(\boldsymbol{\eta}_{0,N}) Y_i \right\} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}) + o_p(|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}|) \quad (\text{A4})$$

en vertu de la continuité uniforme de  $\partial \{\sum_{i \in \mathcal{A}} w_i(\boldsymbol{\eta}) Y_i\} / \partial \boldsymbol{\eta}$  autour de  $\boldsymbol{\eta}_{0,N}$ . Maintenant, en utilisant

$$\frac{\partial}{\partial \boldsymbol{\eta}} g_i(\boldsymbol{\eta}) = - \frac{f(\mathbf{x}_i; \boldsymbol{\eta})}{\partial f(\mathbf{x}_i; \boldsymbol{\eta}) / \partial \boldsymbol{\eta}} \times \frac{f(\mathbf{x}_i; \boldsymbol{\eta})}{f(\mathbf{x}_i; \boldsymbol{\eta})} = -g_i(\boldsymbol{\eta}) \times s_i(\boldsymbol{\eta}),$$

où  $s_i(\boldsymbol{\eta}) = \partial \ln f(\mathbf{x}_i; \boldsymbol{\eta}) / \partial \boldsymbol{\eta}$ , nous avons

$$\frac{\partial}{\partial \boldsymbol{\eta}} \sum_{i \in \mathcal{A}} w_i(\boldsymbol{\eta}) Y_i = - \sum_{i \in \mathcal{A}} w_i(\boldsymbol{\eta}) s_i(\boldsymbol{\eta}) Y_i.$$

**Tableau 2 (suite)**  
Biais Monte Carlo et erreurs quadratiques moyenne Monte Carlo des estimateurs ponctuels de la moyenne de  $y$ , basés sur 10 000 échantillons Monte Carlo

Population	Taille de l'échantillon	Estimateur	Biais		EQM		Biais		EQM	
			EAS	PPT	EAS	PPT	EAS	PPT	EAS	PPT
B	100	Pas de calage	0,00	0,00	0,02044	0,00	0,00	0,00	0,01692	0,00
		Estimateur par régression	-0,01	0,00	0,01473	0,00	0,00	0,00	0,01461	0,00
		Estimateur EL ( $t = 1$ )	0,01	0,00	0,01652	0,01	0,01	0,01	0,01516	0,01
		Estimateur EL ( $t = 10$ )	0,00	0,00	0,01490	0,01	0,01	0,01	0,01472	0,01
		Estimateur ET ( $t = 1$ )	0,00	0,00	0,01516	0,01	0,01	0,01	0,01483	0,01
	200	Estimateur ET ( $t = 10$ )	0,00	0,00	0,01470	0,00	0,00	0,00	0,01459	0,01
		Estimateur IVET ( $t = 1$ )	0,00	0,00	0,01497	0,00	0,00	0,00	0,01458	0,00
		Estimateur IVET ( $t = 10$ )	0,00	0,00	0,01472	0,00	0,00	0,00	0,01453	0,00
		Pas de calage	0,00	0,00	0,00888	0,00	0,00	0,00	0,00823	0,00
		Estimateur par régression	-0,01	0,00	0,00705	0,00	0,00	0,00	0,00735	0,00
	100	Estimateur EL ( $t = 1$ )	0,01	0,00	0,00769	0,01	0,01	0,01	0,00764	0,00
		Estimateur EL ( $t = 10$ )	0,00	0,00	0,00715	0,01	0,01	0,01	0,00745	0,00
		Estimateur ET ( $t = 1$ )	0,00	0,00	0,00723	0,01	0,01	0,01	0,00749	0,00
		Estimateur ET ( $t = 10$ )	0,00	0,00	0,00706	0,01	0,01	0,01	0,00734	0,00
		Estimateur IVET ( $t = 1$ )	0,00	0,00	0,00704	0,00	0,00	0,00	0,00728	0,00
	200	Estimateur IVET ( $t = 10$ )	0,00	0,00	0,00699	0,00	0,00	0,00	0,00725	0,00

EAS, échantillonnage aléatoire simple; PPT, échantillonnage avec probabilité proportionnelle à la taille; EQM, erreur quadratique moyenne; EL, vraisemblance empirique (*empirical likelihood*); ET, inclinaison exponentielle (*exponential tilting*); IVET, inclinaison exponentielle avec variable instrumentale (*instrumental-variable exponential tilting*).

**Tableau 3**  
Biais relatifs Monte Carlo des estimateurs de variance, basés sur 10 000 échantillons Monte Carlo

Population	Taille de l'échantillon	Estimateur	I linéarisation		Rééchantillonnage	
			EAS	PPT	EAS	PPT
A	100	ET ( $t = 1$ )	-7,02	-2,66	10,65	4,11
		ET ( $t = 10$ )	-4,91	-0,80	5,60	0,67
		IVET ( $t = 1$ )	-5,28	-3,63	7,67	2,25
		IVET ( $t = 10$ )	-4,11	-0,87	4,96	0,41
		ET ( $t = 1$ )	-3,97	-0,19	3,65	0,57
	200	ET ( $t = 10$ )	-2,93	0,87	2,23	-0,35
		IVET ( $t = 1$ )	-3,35	-0,10	2,34	0,02
		IVET ( $t = 10$ )	-2,72	0,78	1,62	-0,53
		ET ( $t = 1$ )	-7,64	-3,01	10,72	4,50
		ET ( $t = 10$ )	-5,98	-0,98	7,21	0,74
B	100	IVET ( $t = 1$ )	-5,77	-2,31	4,53	-0,10
		IVET ( $t = 10$ )	-5,44	-1,86	5,17	-0,51
		ET ( $t = 1$ )	-2,41	-1,01	5,76	2,53
		ET ( $t = 10$ )	-1,29	0,18	4,30	1,91
		IVET ( $t = 1$ )	-1,39	-0,35	2,09	1,04
	200	IVET ( $t = 10$ )	-1,15	-0,06	2,04	0,99

EAS, échantillonnage aléatoire simple; PPT, échantillonnage avec probabilité proportionnelle à la taille; ET, inclinaison exponentielle (*exponential tilting*); IVET, inclinaison exponentielle avec variable instrumentale (*instrumental-variable exponential tilting*).

Si (37) possède une solution, celle-ci peut être exprimée comme la limite de la forme

$$w_{it} \propto \prod_{s=1}^{\infty} \exp \left\{ -U' \hat{\Sigma}_{ad(s)}^{-1} U(\mathbf{x}_i) \right\} \quad (38)$$

où  $\hat{U}^{(s)} = \sum_{i \in A} w_{it(s)} U(\mathbf{x}_i)$ ,  $\hat{\Sigma}_{ad(s)} = \sum_{i \in A} w_{it(s)} \{U(\mathbf{x}_i) - \bar{U}^{(s)}\} \{U(\mathbf{x}_i) - \bar{U}^{(s)}\}' / \sum_{i \in A} w_{it(s)}$ . Si la solution de la condition (37) n'existe pas, nous pouvons encore utiliser les poids donnés par (38), mais l'égalité doit être relâchée. À la place,

L'inegalité approximative sera satisfaite dans (37), en ce sens que  $\sum_{i \in A} w_{it(s)} U(\mathbf{x}_i)$  converge vers zéro beaucoup plus rapidement que  $\sum_{i \in A} w_{it(s)} U(\mathbf{x}_i)$  pour  $t \geq 1$ . L'égalité approximative dans (37) est appelée condition de calage approximatif. Les estimateurs  $\hat{Y}^{(t)} = \sum_{i \in A} w_{it(s)} y_i$  dans lesquels sont utilisés les poids ET en  $t$  étapes donnés par (38), y compris l'estimateur en une étape  $\hat{Y}^{(1)}$ , sont asymptotiquement équivalents à l'estimateur par la régression de la forme

$$\hat{Y}^{\text{reg}} = \hat{Y}^{(0)} - U' \hat{\Sigma}_{ad(0)}^{-1} \hat{\Sigma}_{ad(0)} y^{(0)}$$

sous-estime légèrement la variance réelle, parce qu'il omet le terme de deuxième ordre dans la linéarisation de Taylor. L'estimateur de variance par rééchantillonnage présente un léger biais positif dans la simulation. Les biais des estimateurs de variance sont généralement plus faibles en valeur absolue dans la population A, parce que le modèle linéaire est vérifié. Dans la population B, le biais des estimateurs de variance est moins important pour l'estimateur IVET que pour l'estimateur ET, car des poids moins extrême sont utilisés dans l'estimateur IVET.

### 6. Conclusion

Nous avons examiné le problème de l'estimation de  $Y$  en nous servant d'information auxiliaire de la forme  $E\{U(\mathbf{X})\} = 0$  avec une fonction connue  $U(\cdot)$ . Nous avons considéré la classe des estimateurs linéaires de la forme  $\hat{Y} = \sum_{i \in A} w_i Y_i$  avec  $\sum_{i \in A} w_i I_i = U(\mathbf{x}_i) = (\hat{N}_i, 0)$  et  $w_i > 0$ . Si la densité  $f(\mathbf{x}; \boldsymbol{\eta})$  de  $\mathbf{X}$  est connue jusqu'à  $\boldsymbol{\eta} \in \Omega$ , nous pouvons mettre en oeuvre une estimation efficace en utilisant le poids d'échantillonnage préférentiel estimé

$$w_i \propto d_i \frac{f(\mathbf{x}_i; \boldsymbol{\eta}_{0,N})}{f(\mathbf{x}_i; \boldsymbol{\eta})}.$$

où  $d_i$  représente les poids initiaux, et où  $\boldsymbol{\eta}_{0,N}$  et  $\boldsymbol{\eta}$  sont les estimateurs du maximum de vraisemblance de  $\boldsymbol{\eta}$  basé sur la population et sur l'échantillon, respectivement. Si la forme paramétrique de  $f(\mathbf{x}; \boldsymbol{\eta})$  est inconnue, nous pouvons utiliser les poids transformés par inclinaison exponentielle de la forme

$$w_{i(\lambda)} \propto \exp\{\lambda' U(\mathbf{x}_i)\},$$

où  $\lambda$  est déterminé en vue de satisfaire

$$\sum_{i \in A} w_{i(\lambda)} U(\mathbf{x}_i) = 0. \tag{37}$$

En utilisant les échantillons Monte Carlo produits comme il est mentionné plus haut, nous avons calculé les biais et les erreurs quadratiques moyennes de huit estimateurs de la moyenne de population de  $y$ , la variable d'intérêt. Les résultats sont présentés au tableau 2. Les estimateurs par calage contiennent un biais, mais celui-ci est faible si le modèle de régression est vérifié ou si la taille d'échantillon est grande. Dans la population A, le modèle de régression linéaire est vérifié et l'estimateur par la régression est efficace selon les études, celui dont la performance est la meilleure.

En plus de l'estimation ponctuelle, nous avons examiné l'estimation de la variance. Nous n'avons considéré l'estimation de la variance que pour les estimateurs ET et IVET en  $t$  étapes. Nous avons calculé l'estimateur de variance par linéarisation (33) et l'estimateur de variance par rééchantillonnage (36) pour chaque estimateur dans chaque échantillon. Dans la méthode de rééchantillonnage, nous avons utilisé la méthode du jackknife avec suppression d'une unité pour chaque réplique. Nous avons calculé les biais relatifs des estimateurs de variance en divisant le biais Monte Carlo de l'estimateur de variance par la variance Monte Carlo. Les biais relatifs Monte Carlo des estimateurs de variance par linéarisation et des estimateurs de variance par rééchantillonnage sont présentés au tableau 3. Le biais relatif théorique des estimateurs de variance est d'ordre  $o(1)$ , ce qui concorde avec les résultats des simulations présentés au tableau 3. L'estimateur de variance par linéarisation

**Tableau 2**  
Biais Monte Carlo et erreurs quadratiques moyennes Monte Carlo des estimateurs ponctuels de la moyenne de  $y$ , basés sur 10 000 échantillons Monte Carlo

Population	Taille de l'échantillon	Estimateur	EAS	Biais	EQM	Biais	PPT	EQM
A	100	Pas de calage	0,00	0,00	0,02298	0,00	0,00	0,02023
		Estimateur par régression	0,00	0,00	0,01261	0,00	0,00	0,01289
		Estimateur EL ( $t = 1$ )	0,01	0,01	0,01369	0,01	0,01	0,01353
		Estimateur EL ( $t = 10$ )	0,00	0,00	0,01285	0,00	0,00	0,01289
		Estimateur ET ( $t = 1$ )	0,01	0,01	0,01334	0,01	0,01	0,01353
		Estimateur ET ( $t = 10$ )	0,00	0,00	0,01269	0,00	0,00	0,01289
		Estimateur IVET ( $t = 1$ )	0,01	0,01	0,01309	0,01	0,01	0,01330
		Estimateur IVET ( $t = 10$ )	0,00	0,00	0,01263	0,00	0,00	0,01289
	200	Pas de calage	0,00	0,00	0,01069	0,00	0,00	0,00925
		Estimateur par régression	0,00	0,00	0,00595	0,00	0,00	0,00568
		Estimateur EL ( $t = 1$ )	0,01	0,01	0,00632	0,01	0,01	0,00604
		Estimateur EL ( $t = 10$ )	0,00	0,00	0,00597	0,00	0,00	0,00568
		Estimateur ET ( $t = 1$ )	0,00	0,00	0,00616	0,00	0,01	0,00578
		Estimateur ET ( $t = 10$ )	0,00	0,00	0,00596	0,00	0,00	0,00568
		Estimateur IVET ( $t = 1$ )	0,00	0,00	0,00605	0,01	0,01	0,00574
		Estimateur IVET ( $t = 10$ )	0,00	0,00	0,00591	0,00	0,00	0,00567



$$\hat{Y}_{ET}^{(k)} = \sum_{i \in \mathcal{M}} d_i^{(k)} \exp(\hat{\lambda}_{0(1)}^{(k)} + \hat{\lambda}_{1(1)}^{(k)} \mathbf{z}_i^{(k)}) y_i \quad (35)$$

$$\hat{\lambda}_{1(1)}^{(k)} = \left\{ \sum_{i \in \mathcal{M}} d_i^{(k)} (\mathbf{x}_i - \bar{\mathbf{X}}_{(k)}^P) (\mathbf{z}_i - \bar{\mathbf{Z}}_{(k)}^P) / (\hat{N}_{(k)}^P) \right\}^{-1} (\mathbf{X} / \hat{N}_{(k)}^P - \bar{\mathbf{X}}_{(k)}^P),$$

$$\hat{N}_{(k)} = \begin{cases} N & \text{si } \hat{N} = N \\ \sum_{i \in \mathcal{M}} d_i^{(k)} & \text{si } \hat{N} = \hat{N}^P \end{cases}$$

$$(\bar{\mathbf{X}}_{(k)}^P, \bar{\mathbf{Z}}_{(k)}^P) = \left( \sum_{i \in \mathcal{M}} d_i^{(k)} \mathbf{x}_i^{(k)}, \sum_{i \in \mathcal{M}} d_i^{(k)} \mathbf{z}_i^{(k)} \right),$$

et

$$\exp(\hat{\lambda}_{0(1)}^{(k)}) = \frac{\sum_{i \in \mathcal{I}} d_i^{(k)} \exp(\hat{\lambda}_{1(1)}^{(k)})}{\hat{N}}$$

L'estimateur de variance par rééchantillonnage défini par

$$\hat{V}^{\text{rep}} = \sum_{l=1}^k c_l^k (\hat{Y}_{(k)}^{\text{ET}} - \hat{Y}^{\text{ET}})^2, \quad (36)$$

où  $\hat{Y}_{(k)}^{\text{ET}}$  est défini dans (35), peut être utilisé pour estimer la variance de l'estimateur par calage ET donné par (26).

## 5. Étude par simulation

Afin d'étudier les propriétés en échantillon fini des estimateurs proposés, nous avons effectué une étude par simulation limitée. Dans la simulation, nous avons généré indépendamment deux populations finies de taille  $N = 10\,000$ .

Tableau 1  
Exemple de poids de calage avec un échantillon de taille  $n = 5$

Méthode	$\bar{X}_N$	1	2	3	4	5	$\bar{X}_N$
Rég.	3,0	0,200	0,200	0,200	0,200	0,200	3,0
Reg.,	4,5	-0,100	-0,100	0,200	0,035	0,500	4,5
EL	6,0	-0,400	-0,100	0,200	0,500	0,800	6,0
	3,0	0,200	0,200	0,200	0,200	0,200	3,0
	4,5	0,033	0,043	0,063	0,115	0,746	4,5
	6,0	S.O.	S.O.	S.O.	S.O.	S.O.	S.O.
ET ( $t = 1$ )	3,0	0,200	0,200	0,200	0,200	0,200	3,0
	4,5	0,027	0,057	0,100	0,255	0,540	4,2
	6,0	0,002	0,009	0,039	0,173	0,777	4,7
ET ( $t = 10$ )	3,0	0,200	0,200	0,200	0,200	0,200	3,0
	4,5	0,009	0,027	0,078	0,227	0,659	4,5
	6,0	0,000	0,000	0,000	0,001	0,999	5,0
IVET ( $t = 1$ )	3,0	0,200	0,200	0,200	0,200	0,200	3,0
	4,5	0,030	0,047	0,121	0,309	0,493	4,2
	6,0	0,003	0,006	0,041	0,267	0,683	4,6
IVET ( $t = 10$ )	3,0	0,200	0,200	0,200	0,200	0,200	3,0
	4,5	0,007	0,015	0,066	0,294	0,618	4,5
	6,0	0,000	0,000	0,000	0,087	0,913	4,9
Rég., estimateur par la régression ; EL, vraisemblance empirique (empirical likelihood) ; ET, inclinaison exponentielle (exponential tilting) ; IVET, inclinaison exponentielle avec variable instrumentale (instrumental variable exponential tilting) ; S.O., sans objet.							

Pour la population A, la population finie est générée en partant d'une population infinie spécifiée par  $x_i \sim \exp(1) + 1$ ;  $y_i = 3 + x_i + x_i e_i$ ,  $e_i | x_i \sim N(0, 1)$ ;  $z_i | (x_i, y_i) \sim \chi^2(1) + |y_i|$ . Dans le cas de la population B, les  $(x_i, e_i, z_i)$  sont les mêmes que dans la population A, mais  $y_i = (5 - 1/\sqrt{8}) + 1/\sqrt{8}(x_i - 2)^2 + e_i$ . La variable auxiliaire  $x_i$  est utilisée pour le calage et  $z_i$  est la mesure de taille utilisée pour l'échantillonnage avec probabilités inégales. À partir des deux populations finies obtenues, nous avons généré indépendamment  $M = 10\,000$  échantillons Monte Carlo de taille  $n$  sous les deux plans d'échantillonnage décrits plus loin. Le paramètre d'intérêt est la moyenne de population de  $y$  et nous supposons que la taille de population  $N$  est connue. Les conditions de simulation peuvent être décrites comme un plan factoriel  $2 \times 2 \times 8 \times 2$  à quatre facteurs, soit a) deux types de population finie, b) le mécanisme d'échantillonnage : échantillonnage aléatoire simple et échantillonnage avec probabilité proportionnelle à la taille ( $z_i$ ) avec remise, c) la méthode de calage : pas de calage, estimateur par la régression, méthode EL donnée par (6) avec  $t = 1$  et  $t = 10$ , méthode ET en  $t$  étapes donnée par (21) avec  $t = 1$  et  $t = 10$ , et méthode IVET donnée par (26) avec  $t = 1$  et  $t = 10$ , et d) taille de l'échantillon :  $n = 100$  et  $n = 200$ . Puisque l'on suppose que  $N$  est connu, les estimateurs par calage sont calculés de façon à satisfaire  $\sum_{i=1}^n w_i(1, x_i) = (1, \bar{X}_N)$  dans les deux populations. Pour la méthode IVET (26), la variable instrumentale  $z_i$  est créée en utilisant les définitions données en (28) avec le seuil  $C = 3$ .

mêmes estimations de  $\bar{X}_N$  pour les deux valeurs de  $t$ , mais l'estimateur IVET produit des poids moins extrêmes que l'estimateur ET.

#### 4. Estimation de la variance

Examinons maintenant l'estimation de la variance des estimateurs par calage ET des sections 2 et 3. Comme les paramètres estimés  $(\hat{\lambda}_0, \hat{\lambda}_1)$  qui figurent dans l'estimateur par calage ET (16) ont une certaine variabilité d'échantillonnage, la méthode d'estimation de la variance doit en tenir compte. Dans ce cas, l'estimation de la variance peut souvent être obtenue par une méthode de linéarisation ou par une méthode de rééchantillonnage, ou réplique (Wolter 2007). Pour la discussion de la méthode de linéarisation, posons que la variance de l'estimateur HT donné en (1) est estimée de manière convergente par

$$V(Y^d) = \sum_{i \in A} \sum_{j \in J} \Omega_{ij} y_i y_j. \quad (32)$$

Pour l'estimateur ET, l'estimateur de variance par linéarisation peut être obtenu au moyen de la formule d'estimation de la variance par linéarisation établie pour l'estimateur par la régression, comme dans Deville et Särndal (1992), en se servant de l'équivalence asymptotique entre l'estimateur par calage ET et l'estimateur par la régression montrée dans le théorème 2. En particulier, si l'on connaît la taille de population  $N$ , un estimateur de variance par linéarisation pour l'estimateur IVET (26) peut s'écrire

$$V(Y^{\text{IVET}}) = \sum_{i \in A} \sum_{j \in J} \Omega_{ij} g_i g_j \hat{e}_i \hat{e}_j \quad (33)$$

où  $\Omega_{ij}$  correspond aux coefficients de l'estimateur de variance (32),  $g_i = w_i/d_i$  est le facteur d'ajustement des poids et  $\hat{e}_i = y_i - \bar{Y}^d - (\mathbf{x}_i - \bar{\mathbf{X}}^d)' \mathbf{B}^d$ , où  $\mathbf{B}^d$  est défini dans (30). Le choix  $\mathbf{z}_i = \mathbf{x}_i$  dans (33) produit l'estimateur de variance linéarisé pour l'estimateur ET donné par (16). La démonstration de la convergence de l'estimateur de variance (33) peut être consultée dans Kim et Park (2010). Pour l'estimateur ET en une étape, il est facile de mettre en œuvre une méthode de rééchantillonnage (réplique). Soit l'estimateur de variance par rééchantillonnage de la

forme

$$V^{\text{rep}} = \sum_{k=1}^L c_k (Y_{(k)}^d - \bar{Y}^d)^2, \quad (34)$$

où  $L$  est le nombre de répliques, et  $c_k$  est le facteur de réplique associé à la réplique  $k$ ,  $Y_{(k)}^d = \sum_{i \in A} d_i^{(k)} y_i$ , et  $d_i^{(k)}$  est la  $k^{\text{e}}$  réplique du poids d'échantillonnage  $d_i$ . Par exemple, l'estimateur de variance par rééchantillonnage (34) inclut le jackknife et le bootstrap (voir Rust et Rao 1996). Supposons que l'estimateur de variance par rééchantillonnage (34) est un estimateur convergent pour la variance de  $Y^d$ . La  $k^{\text{e}}$  réplique de l'estimateur ET en une étape peut être calculée par

$$w_i = d_i g_i(\hat{\lambda}) = d_i \frac{L(U - 1) + U(1 - L) \exp(K \hat{\lambda} \mathbf{x}_i)}{(U - 1) + (1 - L) \exp(K \hat{\lambda} \mathbf{x}_i)}. \quad (31)$$

où  $K = (U - L) / \{(1 - L)(U - 1)\}$ , pour des valeurs de  $L$  et  $U$  telles que  $0 < L < 1 < U$ . Si les contraintes de calage (2) et (4) doivent être satisfaites, nous pouvons utiliser  $\hat{\lambda}_0 + \hat{\lambda}_1' \mathbf{x}_i$  au lieu de  $\hat{\lambda} \mathbf{x}_i$  dans (31). L'estimateur par calage résultant est asymptotiquement équivalent à l'estimateur par la régression avec utilisation des poids donnés par (5), tandis que l'estimateur IVET est asymptotiquement équivalent à l'estimateur par la régression à variable instrumentale (29). Les calculs en vue d'obtenir  $\hat{\lambda}$  sont assez compliqués, parce que  $\partial g_i(\hat{\lambda}) / \partial \hat{\lambda}$  n'est pas facile à évaluer dans (31). Dans l'estimateur IVET, le calcul, donné par (27), est simple.

Afin de comparer la pondération proposée aux méthodes existantes, considérons l'exemple artificiel d'un échantillon aléatoire simple de taille  $n = 5$  où  $x_k = k$ ,  $k = 1, 2, \dots, 5$ . Nous exécutons les calculs pour trois moyennes de population de  $x$ ;  $\bar{X}_N = 3$ ,  $\bar{X}_N = 4.5$ , et  $\bar{X}_N = 6$ . Le tableau 1 donne les poids résultants pour l'estimateur par la régression, l'estimateur par la vraisemblance empirique (EL), l'estimateur par inclinaison exponentielle (ET) en  $t$  étapes (16) avec  $t = 1$  et  $t = 10$ , ainsi que l'estimateur par inclinaison exponentielle à variable instrumentale (IVET) en  $t$  étapes (26) avec  $t = 1$  et  $t = 10$ . Pour l'estimateur IVET, nous créons la variable instrumentale  $z_i$  telle que

$$z_i = \begin{cases} 1.5 & \text{si } x_i \leq 1.5 \\ x_i & \text{si } x_i \in (1.5; 4.5) \\ 4.5 & \text{si } x_i \geq 4.5. \end{cases}$$

La dernière colonne du tableau 1 donne la moyenne estimée de  $X$  en utilisant les poids de calage respectifs. Tous les poids sont égaux à  $1/n = 0.2$  pour  $\bar{X}_N = 3$ . L'estimateur par la régression est linéairement croissant en  $x_i$ , mais possède des poids négatifs pour les populations de moyenne  $\bar{X}_N = 4.5$  et de moyenne  $\bar{X}_N = 6$ . Pour la population de moyenne  $\bar{X}_N = 6$ , les poids  $n$  ont pas pu être calculés pour la méthode EL, parce que  $\bar{X}_N$  se situe en dehors de l'intervalle des valeurs  $x_i$  d'échantillon. Dans ce cas extrême où  $\bar{X}_N = 6$ , la méthode ET fournit des poids non négatifs en sacrifiant la contrainte de calage et l'estimateur EL possède des poids plus extrêmes que l'estimateur ET ou que l'estimateur IVET en ce sens que le poids pour  $k = 5$  est le plus grand observé parmi les estimateurs étudiés. Le poids pour l'estimateur ET en une étape est proche de celui de l'estimateur par la régression pour une grande valeur de  $x_i$ , mais il est proche de celui de l'estimateur EL pour une petite valeur de  $x_i$ . Les estimateurs ET en 10 étapes ont de meilleures propriétés de calage en ce sens que l'erreur quadratique,  $(\sum_{k=1}^5 w_k x_k - \bar{X}_N)^2$ , est plus faible que pour l'estimateur ET en une étape. L'estimateur ET et l'estimateur IVET donnent presque les

$$\mathcal{Q}(w) = \sum_{i \in A} w_i \ln \left( \frac{d_i}{w_i} \right) \quad (25)$$

sous les contraintes (2) et (4). La fonction objectif (25) est souvent appelée fonction de discrimination minimale. La valeur minimale de  $\mathcal{Q}(w)$  est zéro si (4) est la seule contrainte de calage et elle croît de manière monotone si des contraintes de calage additionnelles sont imposées.

### 3. Calage au moyen de variables instrumentales

Nous considérons une extension de la méthode proposée à la section 2 à une classe plus générale d'estimateurs par calage ET utilisant des variables instrumentales. Estevao et Särndal (2000) ainsi que Kott (2003) ont discuté de l'utilisation de variables instrumentales pour l'estimation par calage dans le contexte de simulations limitées. Soit  $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i)$  une variable instrumentale dérivée de  $\mathbf{x}_i$ , où la fonction  $\mathbf{z}(\cdot)$  doit être déterminée. L'estimateur par inclusion exponentielle avec variable instrumentale (IVET) (pour *instrumental-variable exponential tilting*) en se servant de la variable instrumentale  $\mathbf{z}_i$  peut être défini comme

$$\hat{Y}^{\text{IVET}} = \sum_{i \in I} w_i y_i = \sum_{i \in I} d_i \exp(\hat{\lambda}_0 + \hat{\lambda}_i' \mathbf{z}_i) y_i, \quad (26)$$

où  $\hat{\lambda}_0$  et  $\hat{\lambda}_i$  sont calculés d'après (2) et (4). Notons que l'estimateur IVET donné par (26) appartient à une classe d'estimateurs indexés par  $\mathbf{z}_i$ . L'approche de la variable instrumentale définie en (26) offre plus de souplesse pour créer l'estimateur ET. Le choix de  $\mathbf{z}_i = \mathbf{x}_i$  aboutit à l'estimateur ET standard donné par (16), mais une certaine transformation  $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i)$  peut rendre l'estimateur ET donné par (26) plus intéressant en pratique. La solution des équations de calage peut être obtenue itérativement comme il suit

$$\hat{\lambda}_{i(t+1)} = \hat{\lambda}_{i(t)} + \left\{ \sum_{i \in A} w_{i(t)} (\mathbf{x}_i - \bar{\mathbf{X}} - \bar{\mathbf{X}}^{w(i)})(\mathbf{z}_i - \bar{\mathbf{Z}} - \bar{\mathbf{Z}}^{w(i)}) \right\}^{-1} \left( \mathbf{X} - \sum_{i \in A} w_{i(t)} \mathbf{x}_i \right), \quad (27)$$

où  $w_{i(t)} = d_i \exp(\hat{\lambda}_{0(t)} + \hat{\lambda}_{i(t)}' \mathbf{z}_i)$  et  $\bar{\mathbf{Z}}^{w(i)} = \sum_{i \in A} w_{i(t)} \mathbf{z}_i / \sum_{i \in A} w_{i(t)}$ , avec l'équation (20) inchangée et  $\hat{\lambda}_{i(0)} = \mathbf{0}$ .

L'estimateur IVET (26) est utilisé pour créer des poids finaux ayant des valeurs moins extrêmes. Puisque dans (26), le poids final est une fonction de  $\mathbf{z}_i$ , nous pouvons borner  $\mathbf{g}_i = w_i / d_i$  en donnant des bornes à  $\mathbf{z}_i$ . Pour créer la variable  $\mathbf{z}_i$  bornée, nous pouvons utiliser une version tronquée de  $\mathbf{x}_i$ , désignée par  $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{ip})$ , où

$$z_{ij} = \begin{cases} \bar{x}_j & \text{si } |x_{ij} - \bar{x}_j| \leq C_j S_j \\ \bar{x}_j + C_j S_j & \text{si } x_{ij} > \bar{x}_j + C_j S_j \\ \bar{x}_j - C_j S_j & \text{si } x_{ij} < \bar{x}_j - C_j S_j \end{cases} \quad (28)$$

forme

Remarque 3. Deville et Särndal (1992) ont également considéré des poids de calage à étendue restreinte de la forme

$$\hat{\mathbf{B}}_z = \left\{ \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_{(i)}) \Phi_i (\mathbf{x}_i - \bar{\mathbf{X}}_{(i)})' \right\}^{-1} \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_{(i)}) \Phi_i (\mathbf{x}_i - \bar{\mathbf{X}}_{(i)})',$$

et où  $\hat{\mathbf{B}}_z = \left\{ \sum_{i \in A} d_i (\mathbf{z}_i - \bar{\mathbf{Z}}_{(i)}) (\mathbf{x}_i - \bar{\mathbf{X}}_{(i)}) (\mathbf{z}_i - \bar{\mathbf{Z}}_{(i)})' \right\}^{-1} \sum_{i \in A} d_i (\mathbf{z}_i - \bar{\mathbf{Z}}_{(i)}) (\mathbf{x}_i - \bar{\mathbf{X}}_{(i)}) (\mathbf{z}_i - \bar{\mathbf{Z}}_{(i)})' y_i$ . (30)

L'estimateur (29) prend la forme d'un estimateur par la régression à variable instrumentale. Donc, sous le choix  $\mathbf{z}_i = \Phi_i' \mathbf{x}_i$ , l'estimateur par la régression à variable instrumentale peut s'écrire comme (29) avec

$$\hat{Y}_{\text{IV,reg}} = \tilde{Y}_d + (\mathbf{X} - \tilde{\mathbf{X}}_d)' \hat{\mathbf{B}}_z \quad (29)$$

valent à

Au lieu d'employer la variable instrumentale tronquée  $\mathbf{z}_i$  dans (28), nous pouvons considérer la variable instrumentale suivante

$$\mathbf{z}_i = \mathbf{x}_i \Phi_i$$

utilisé comme alternative à la troncature des poids. Au lieu d'employer la variable instrumentale tronquée  $\mathbf{z}_i$  dans (28), nous pouvons considérer la variable instrumentale suivante



où  $\hat{\lambda}_0$  et  $\hat{\lambda}_1$  satisfont les contraintes (2) et (4). Les estimateurs basés sur l'inclinaison exponentielle ont été utilisés dans divers contextes. Consulter, par exemple, Efron (1981), Kitamura et Stutzer (1997), et Imbens (2002). Pour  $N$  connu, Folsom (1991) ainsi que Deville, Särndal et Sautory (1993) ont élaboré l'estimateur (16) en utilisant une approche fort différente.

Afin de calculer  $\hat{\lambda}_0$  et  $\hat{\lambda}_1$  dans (16), à cause des contraintes de calage (2) et (4), nous devons résoudre les estimations suivantes :

$$\hat{U}_0(\lambda) \equiv \sum_{i \in A} d_i \exp(\lambda_0 + \lambda_1' \mathbf{x}_i) - N = 0 \quad (17)$$

$$\hat{U}_1(\lambda) \equiv \sum_{i \in A} d_i \exp(\lambda_0 + \lambda_1' \mathbf{x}_i) \mathbf{x}_i - \mathbf{X} = \mathbf{0}, \quad (18)$$

où  $\lambda' = (\lambda_0, \lambda_1')$ . En écrivant  $\hat{\mathbf{U}}' = (\hat{U}_0, \hat{U}_1)$ , nous pouvons utiliser l'algorithme de type Newton de la forme

$$\hat{\lambda}^{(t+1)} = \hat{\lambda}^{(t)} - \left\{ \frac{\partial}{\partial \lambda} \hat{\mathbf{U}}(\hat{\lambda}^{(t)}) \right\}^{-1} \hat{\mathbf{U}}(\hat{\lambda}^{(t)})$$

et la solution peut s'écrire

$$\hat{\lambda}_{1(t+1)} = \hat{\lambda}_{1(t)}$$

$$+ \left\{ \sum_{i \in A} w_{i(t)} (\mathbf{x}_i - \bar{\mathbf{X}}^{w(t)}) \otimes_2^{-1} \left( \mathbf{X} - \sum_{i \in A} w_{i(t)} \mathbf{x}_i \right) \right\}, \quad (19)$$

où  $w_{i(t)} = d_i \exp(\hat{\lambda}_{0(t)} + \hat{\lambda}_{1(t)}' \mathbf{x}_i)$  et  $\bar{\mathbf{X}}^{w(t)} = \sum_{i \in A} w_{i(t)} \mathbf{x}_i / \sum_{i \in A} w_{i(t)}$ , avec les valeurs initiales  $\hat{\lambda}_{1(0)} = \mathbf{0}$ . Après avoir calculé  $\hat{\lambda}_{1(t)}$ , en nous servant de (19), nous calculons  $\hat{\lambda}_{0(t)}$  comme il suit

$$\exp(\hat{\lambda}_{0(t)}) = \frac{\sum_{i \in A} d_i \exp(\hat{\lambda}_{1(t)}' \mathbf{x}_i)}{N}. \quad (20)$$

Notons que  $w_{i(0)} = d_i N / N^d$  car  $\hat{\lambda}_{1(0)} = \mathbf{0}$ .  $\hat{\mathbf{U}}(\lambda)$  étant deux fois continuellement dérivable et convexe en  $\lambda$ , la série  $\hat{\lambda}^{(t)}$  converge toujours si la solution de  $\hat{\mathbf{U}}(\lambda) = \mathbf{0}$  existe (Givens et Hoeting 2005). Le taux de convergence est quadratique en ce sens que

$$|\hat{\lambda}_{1(t+1)} - \hat{\lambda}_1| \leq C |\hat{\lambda}_{1(t)} - \hat{\lambda}_1|^2$$

pour une constante  $C$ , où  $\hat{\lambda}_1 = \lim_{t \rightarrow \infty} \hat{\lambda}_{1(t)}$ . Par construction, l'estimateur par inclinaison exponentielle (ET) en  $t$  étapes, défini par

$$\hat{Y}_{\text{ET}(t)} = \sum_{i \in A} d_i \exp(\hat{\lambda}_{0(t)} + \hat{\lambda}_{1(t)}' \mathbf{x}_i) y_i \quad (21)$$

où  $\hat{\lambda}_{0(t)}$  et  $\hat{\lambda}_{1(t)}$  sont calculés au moyen de (19) et (20), satisfait la contrainte de calage (2) pour une valeur

suffisamment grande de  $t$ . En vertu de la forme récursive dans (19) avec  $\hat{\lambda}_{1(0)} = \mathbf{0}$ , nous pouvons écrire

$$\hat{\lambda}_{1(t)} = \sum_{l=0}^{t-1} (\mathbf{S}^{\text{rw}(l)})^{-1} (\bar{\mathbf{X}}^N - \bar{\mathbf{X}}^{w(l)}), \quad (22)$$

où  $\bar{\mathbf{X}}^N = \mathbf{X}/N$  et  $\mathbf{S}^{\text{rw}(j)} = \sum_{i \in A} w_{i(j)} (\mathbf{x}_i - \bar{\mathbf{X}}^{w(j)})^{\otimes 2} / \hat{N}$ . Donc, l'estimateur ET en  $t$  étapes (21) peut s'écrire

$$\hat{Y}_{\text{ET}(t)} = N \frac{\sum_{i \in A} d_i g_{it(i)}}{\sum_{i \in A} d_i g_{it(i)} y_i},$$

où

$$g_{it(i)} = \frac{\prod_{l=1}^t \phi(\mathbf{x}_i; \bar{\mathbf{X}}^N, \mathbf{S}^{\text{rw}(l)})}{\prod_{l=1}^t \phi(\mathbf{x}_l; \bar{\mathbf{X}}^{w(l)}, \mathbf{S}^{\text{rw}(l)})}.$$

Le théorème qui suit présente certaines propriétés asymptotiques de l'estimateur ET en  $t$  étapes (21) basé sur les équations (19) et (20) satisfait

$$\sqrt{n} N^{-1} (Y_{\text{ET}(t)} - Y_{\text{reg}}) = o_p(1), \quad (23)$$

pour chaque  $t = 1, 2, \dots$ , où  $Y_{\text{reg}}$  est l'estimateur par la régression en utilisant les poids de régression (5).

La preuve du théorème 2 est présentée à l'annexe B. Le théorème 2 donne l'équivalence asymptotique entre l'estimateur ET en  $t$  étapes et l'estimateur par la régression. Contrairement à ce dernier, les poids de l'estimateur ET sont toujours positifs. Pour une valeur suffisamment grande de  $t$ , l'estimateur ET en  $t$  étapes satisfait la contrainte de calage (2). Deville et Särndal (1992) ont prouvé le résultat (23) pour le cas particulier où  $t \rightarrow \infty$ .

**Remarque 1.** L'estimateur ET en une étape, défini par  $Y_{\text{ET}(1)}$ , possède un paramètre d'inclinaison ayant une expression analytique

$$\hat{\lambda}_{1(1)} = \left\{ \sum_{i \in A} d_i' (\mathbf{x}_i - \bar{\mathbf{X}}^d)' \right\} / \hat{N}^d \left( \bar{\mathbf{X}}^N - \bar{\mathbf{X}}^d \right), \quad (24)$$

où  $\bar{\mathbf{X}}^N = \mathbf{X}/N$  et  $\bar{\mathbf{X}}^d = \sum_{i \in A} d_i' \mathbf{x}_i / \sum_{i \in A} d_i'$ . En vertu du théorème 2, l'estimateur ET en une étape est asymptotiquement équivalent à l'estimateur par la régression, mais la contrainte de calage (2) n'est pas nécessairement satisfaite. En utilisant le théorème 2 appliqué à  $\mathbf{x}_i$  au lieu de  $y_i$ , on peut montrer que l'estimateur ET en une étape satisfait la contrainte de calage approximativement (Remarque 2. L'estimateur ET peut également être dérivé en trouvant les poids qui minimisent

et soit  $\mathbf{y}_{0N}$  l'estimateur du maximum de vraisemblance de  $\mathbf{y}$  calculé d'après la population

$$\mathbf{y}_{0N} = \arg \max_{\mathbf{y}} \sum_{i=1}^I \ln \{f(\mathbf{x}_i; \mathbf{y})\}.$$

En nous inspirant de Henmi et coll. (2007), nous pouvons construire le poids d'échantillonnage préférentiel estimé

suisant

$$w_i = d_i f(\mathbf{x}_i; \mathbf{y}_{0N}) / f(\mathbf{x}_i; \mathbf{y}). \quad (8)$$

Afin de discuter des propriétés asymptotiques de l'estimateur qui utilise les poids donnés par (8), supposons que nous avons une série de populations finies et d'échantillons, comme dans Isaki et Fuller (1982), tels que

$$\sum_{i=1}^I d_i (\mathbf{x}'_i, \mathbf{y}'_i)' (\mathbf{x}'_i, \mathbf{y}_i)' - \sum_{i=1}^I (\mathbf{x}'_i, \mathbf{y}'_i)' (\mathbf{x}'_i, \mathbf{y}_i)' = O_p(n^{-1/2} N)$$

pour tout  $A$  possible et pour chaque  $N$ . Le théorème qui suit donne certaines propriétés asymptotiques de l'estimateur avec les poids d'échantillonnage préférentiel estimés

par (8), satisfait

$$\sqrt{n} N^{-1} (\hat{Y}_w - Y) = o_p(1), \quad (9)$$

où

$$\hat{Y}_w = \hat{Y} - \hat{\Sigma}^{-1}_{sv} \hat{\Sigma}^{-1}_{ss} \hat{S}_{0d}, \quad (10)$$

$\hat{Y}_d$  est défini en (1),  $\hat{S}_{0d} = \sum_{i \in A} d_i \mathbf{s}_{i0}$ ,  $\hat{\Sigma}_{sv} = N^{-1} \sum_{i \in A} d_i \mathbf{s}_{i0} \mathbf{y}_i'$  et  $\hat{\Sigma}_{ss} = N^{-1} \sum_{i \in A} d_i \mathbf{s}_{i0} \mathbf{s}_{i0}'$ . Ici,  $\mathbf{s}_{i0} = \partial \ln f(\mathbf{x}_i; \mathbf{y}) / \partial \mathbf{y}|_{\mathbf{y}=\mathbf{y}_{0N}}$  la notation  $B^{\otimes 2}$  désigne  $B B'$ .

La preuve du théorème 1 est présentée à l'annexe A.

Comme  $\mathbf{S}_{0N} \equiv \sum_{i=1}^I \mathbf{s}_{i0} = \mathbf{0}$ , nous pouvons écrire (10) sous la forme

$$\hat{Y}_l = \hat{Y} + \hat{\Sigma}^{-1}_{sv} \hat{\Sigma}^{-1}_{ss} (\mathbf{S}_{0N} - \hat{\mathbf{S}}_{0d}),$$

qui est un estimateur par la régression de  $Y$  utilisant  $\mathbf{s}_i(\mathbf{y}_{0N})$  comme variable auxiliaire. Par conséquent, sous des conditions de régularité, l'estimateur proposé utilisant l'échantillonnage préférentiel avec distribution estimée est asymptotiquement sans biais et possède une variance directe  $\hat{Y}_d$ . Notons que la validité du théorème 1 ne requiert pas que le modèle de travail  $f(\mathbf{x}; \mathbf{y})$  soit vrai. Si la densité de  $\mathbf{x}_i$  est une densité normale multivariée, alors les poids donnés par (8) deviennent

$$w_i = d_i \frac{\phi(\mathbf{x}_i; \underline{\mathbf{X}}_{N^*}, \underline{\Sigma}^{xx, d})}{\phi(\mathbf{x}_i; \underline{\mathbf{X}}_{N^*}, \underline{\Sigma}^{xx, d})}, \quad (11)$$

où  $\underline{\mathbf{X}}^d$  est défini comme dans (5),  $\underline{\Sigma}^{xx, d} = \sum_{i \in A} d_i (\mathbf{x}_i - \bar{\mathbf{x}}^d)(\mathbf{x}_i - \bar{\mathbf{x}}^d)'/N^d$ ,  $\underline{\Sigma}^{xx, N} = \sum_{i=1}^I (\mathbf{x}_i - \bar{\mathbf{x}}^N)(\mathbf{x}_i - \bar{\mathbf{x}}^N)'/N$  et  $\phi(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$  est la densité de la loi normale multivariée de moyenne  $\boldsymbol{\mu}$  et de matrice de variance-covariance  $\Sigma$ . Si  $\underline{\Sigma}^{xx, N}$  est inconnue et que seul  $\underline{\mathbf{X}}^N$  est disponible, nous pouvons utiliser

$$w_i = d_i \frac{\phi(\mathbf{x}_i; \underline{\mathbf{X}}_{N^*}, \underline{\Sigma}^{xx, d})}{\phi(\mathbf{x}_i; \underline{\mathbf{X}}_{N^*}, \underline{\Sigma}^{xx, d})}. \quad (12)$$

Tillé (1998) a dérivé des poids similaires à ceux donnés par (12) dans le contexte des probabilités d'inclusion conditionnelles.

En général, le modèle paramétrique pour  $\mathbf{x}_i$  est inconnu. Donc, nous considérons une approximation des poids d'échantillonnage préférentiel donnés par (8) en utilisant le critère d'information de Kullback-Leibler comme distance. Soit  $f(\mathbf{x})$  une densité donnée pour  $\mathbf{x}$  et soit  $P_0$  l'ensemble des densités qui satisfont la contrainte de calage. Autrement dit,

$$P_0 = \left\{ f_0(\mathbf{x}); \int f_0(\mathbf{x}) d\mathbf{x} = 1, \int \mathbf{x} f_0(\mathbf{x}) d\mathbf{x} = \underline{\mathbf{X}}^N \right\}.$$

Le problème d'optimisation en utilisant la distance de Kullback-Leibler peut s'exprimer sous la forme

$$\min_{f_0 \in P_0} \int f_0(\mathbf{x}) \ln \left\{ \frac{f(\mathbf{x})}{f_0(\mathbf{x})} \right\} d\mathbf{x}. \quad (13)$$

La solution de (13) est

$$f_0(\mathbf{x}) = f(\mathbf{x}) \frac{\exp(\lambda \mathbf{x})}{\exp(\lambda \hat{\mathbf{x}})} \quad (14)$$

où  $\hat{\lambda}$  satisfait  $\int \mathbf{x} f_0(\mathbf{x}) d\mathbf{x} = \underline{\mathbf{X}}^N$ . Donc, les poids d'échantillonnage préférentiel estimés donnés par (8) peuvent s'écrire, en utilisant la densité optimale (14), sous la forme

$$w_i = d_i \frac{f_0(\mathbf{x}_i)}{f(\mathbf{x}_i)} = d_i \exp(\hat{\lambda}_0 + \hat{\lambda}_1' \mathbf{x}_i) \quad (15)$$

où  $\hat{\lambda}_0$  et  $\hat{\lambda}_1$  satisfont les contraintes (2) et (4). Le déplacement de  $f(\mathbf{x})$  et de  $f_0(\mathbf{x})$  dans (14) est appelé inclinaison exponentielle (*exponential tilting*). Donc, un estimateur utilisant le poids (15) satisfaisant les contraintes de calage exponentielle (ET) pour *exponential tilting*. Autre-ment dit, nous définissons l'estimateur par calage ET comme il suit

$$Y_{ET} = \sum_{i \in A} d_i \exp(\hat{\lambda}_0 + \hat{\lambda}_1' \mathbf{x}_i) y_i, \quad (16)$$

où  $\hat{\mathbf{X}}_d = \sum_{i \in A} d_i \mathbf{x}_i$ ,  $\hat{N}_d = \sum_{i \in A} d_i$  et  $\hat{\mathbf{X}}_d = \hat{\mathbf{X}}_d / \hat{N}_d$ . Nous définissons l'estimateur par la régression comme étant  $\mathbf{X}_i^{\text{reg}} = \sum_{i \in A} w_i y_i$  en utilisant les poids (5). L'estimateur par la régression peut être efficace si  $y_i$  est relié linéairement à  $\mathbf{x}_i$  (Isaki et Fuller 1982 ; Fuller 2002), mais les poids dans cet estimateur peuvent prendre des valeurs négatives ou extrêmement grandes.

L'estimateur par calage par la vraisemblance empirique (EL pour *empirical likelihood*) dont discutent Chen et Qin (1993), Chen et Sitter (1999), Wu et Rao (2006), et Kim (2009) s'obtient en maximisant la pseudo-vraisemblance empirique

$$\sum_{i \in A} d_i \ln(w_i)$$

sous les contraintes (2) et (4). La solution du problème d'optimisation peut s'écrire

$$w_i = d_i \frac{\lambda_0 + \lambda_1'(\mathbf{x}_i - \mathbf{X}/N)}{1} \quad (6)$$

où  $\lambda_0$  et  $\lambda_1$  satisfont les contraintes (2), (4) et  $w_i > 0$  pour tout  $i$ . L'estimateur par calage EL est asymptotiquement équivalent à l'estimateur par la régression avec utilisation des poids (5) et évite l'obtention de poids négatifs si une solution existe, mais peut produire des poids extrêmement

grands.

Comme la méthode de la vraisemblance empirique requiert la résolution d'équations non linéaires, les calculs peuvent être fastidieux. En outre, dans certains cas extrêmes,  $\hat{\mathbf{X}} = N^{-1} \sum_{i=1}^N \mathbf{x}_i$  n'appartient pas à l'enveloppe convexe des  $\mathbf{x}_i$  d'échantillon et la solution n'existe pas. Le cas échéant, la contrainte (2) peut être relâchée.

Rao et Singh (1997) ont résolu un problème similaire en permettant que

$$\left| \sum_{i \in A} w_i x_{ij} - X_j \right| \leq \delta_j X_j, \quad j = 1, 2, \dots, p,$$

pour un seuil de tolérance donné faible  $\delta_j > 0$ , où  $X_j = \sum_{i=1}^N x_{ij}$ . Notons que le choix  $\delta_j = 0$  aboutit à la condition de calage exact (2). Rao et Singh (1997) ont choisi le seuil de tolérance  $\delta_j$  en utilisant un facteur de rétrécissement dans la régression ridge, mais leur approche ne s'applique pas directement à la méthode de la vraisemblance empirique et le choix de  $\delta_j$  n'est pas tout à fait clair.

Chambers (1996), et Beaumont et Bocci (2008) ont également discuté de l'estimation par la régression ridge comme moyen d'éviter les poids extrêmes. Breidt, Claeskens et Opsomer (2005) ont suivi l'approche des splines pénalisées pour obtenir le calage ridge. Récemment, Park et Fuller (2009) ont élaboré une méthode d'obtention du facteur de rétrécissement  $\delta_j$  en utilisant un modèle de régression en superpopulation avec composantes aléatoires.

Chen, Varaiya et Abraham (2008) ont essayé de résoudre un problème semblable dans le contexte de la méthode du maximum de vraisemblance empirique et ont proposé une solution en ajoutant un point artificiel, tel que  $\hat{\mathbf{X}} = N^{-1} \sum_{i=1}^N \mathbf{x}_i$  appartenirait à l'enveloppe convexe des indices  $\mathbf{x}_i$  augmentés. L'estimateur proposé dans Chen et coll. (2008) ne satisfait la propriété de calage qu'approximativement en ce sens que

$$\sum_{i \in A} w_i \mathbf{x}_i - \mathbf{X} = o_p(n^{-1/2} N). \quad (7)$$

Cette propriété de calage approximatif est intéressante, parce qu'elle permet une plus grande généralité dans le choix des poids. En particulier, quand la dimension de la variable auxiliaire  $\mathbf{x}$  est grande, la contrainte de calage (2) peut être assez restreignante. Comme nous le montrons à la section 2, un estimateur satisfaisant la propriété de calage asymptotique (7) possède la plupart des propriétés désirables de l'estimateur par calage par la vraisemblance empirique et est efficace sur le plan des calculs.

Par le présent article, nous considérons une classe d'estimateurs de type vraisemblance empirique qui satisfont la propriété de calage approximatif (7). À la section 2, nous discutons de l'idée de l'échantillonnage préférentiel avec distribution d'échantillonnage estimée de Hemmi et coll. (2007), et proposons un nouvel estimateur s'appuyant sur cette méthode. À la section 3, nous proposons une technique de troncature des poids pour éviter les poids de calage extrêmes. À la section 4, nous discutons de l'estimation de la variance de l'estimateur proposé. À la section 5, nous exposons les résultats d'une étude par simulation. Enfin, à la section 6, nous présentons nos conclusions.

## 2. Méthode proposée

Avant de présenter la méthode que nous proposons, nous discutons de l'échantillonnage préférentiel introduit par Hemmi et coll. (2007). Supposons que  $\mathbf{x}_i$  est observé dans toute la population, mais que  $y_i$  est observé uniquement dans l'échantillon. Nous émettons l'hypothèse d'un modèle de superpopulation pour  $\mathbf{x}_i$  dont la densité  $f(\mathbf{x}; \boldsymbol{\eta})$  est connue jusqu'à un paramètre  $\boldsymbol{\eta} \in \Omega$ . Le modèle de superpopulation caractérisé par la densité  $f(\mathbf{x}; \boldsymbol{\eta})$  est de travail en ce sens qu'il est utilisé pour calculer un estimateur assisté par modèle (Särndal, Swenson et Wretman 1992).

Soit  $\hat{\boldsymbol{\eta}}$  l'estimateur du pseudo-maximum de vraisemblance de  $\boldsymbol{\eta}$  calculé d'après l'échantillon

$$\hat{\boldsymbol{\eta}} = \arg \max_{\boldsymbol{\eta} \in \Omega} \sum_{i \in A} d_i \ln \{f(\mathbf{x}_i; \boldsymbol{\eta})\}$$



# Estimation par calage en utilisant l'inclinaison exponentielle dans les enquêtes par sondage

Jae Kwang Kim<sup>1</sup>

## Résumé

Nous considérons le problème de l'estimation des paramètres au moyen d'information auxiliaire, quand celle-ci prend la forme de moments connus. L'estimation par calage est un exemple type de l'utilisation des conditions des moments dans les enquêtes par sondage. Étant donné la forme paramétrique de la distribution originale des observations de l'échantillon, nous utilisons l'échantillonnage préférentiel avec distribution d'échantillonnage estimée de Henmi, Yoshida et Eguchi (2007) pour obtenir un estimateur amélioré. Si nous nous servons de la densité normale pour calculer les poids d'échantillonnage préférentiel, l'estimateur résultant prend la forme d'un estimateur par inclinaison exponentielle en une étape. Nous montrons que l'estimateur par inclinaison exponentielle proposé est asymptotiquement équivalent à l'estimateur par la régression, mais qu'il permet d'éviter les poids extrêmes et offre des avantages du point de vue des calculs par rapport à l'estimateur de la vraisemblance empirique. Nous discutons également de l'estimation de la variance et présentons les résultats d'une étude par simulation limitée.

Mots clés : Estimateur par étalonnage ; vraisemblance empirique ; calage au moyen de variables instrumentales ; échantillonnage préférentiel ; estimateur par la régression.

## 1. Introduction

L'estimateur par la régression, en utilisant les poids

$$w_i = d_i + (\mathbf{X} - \hat{\mathbf{X}})' \sum_{j \in A} d_j' \mathbf{x}_j' \quad (3)$$

obtenus en minimisant

$$\sum_{i \in A} (w_i - d_i)^2 / d_i$$

sous la contrainte (2), est asymptotiquement sans biais par rapport au plan. Notons que, si un terme constant est inclus dans l'espace colonne de la matrice  $\mathbf{X}$ , alors (2) implique que la taille de population  $N$  est connue. Si  $N$  est inconnu, on peut exiger que la somme des poids finaux soit égale à la somme des poids d'échantillonnage. Donc,

$$\sum_{i \in A} w_i = N, \quad (4)$$

où

$$\hat{N} = \begin{cases} N & \text{si } N \text{ est connu} \\ \sum_{i \in A} d_i & \text{autrement,} \end{cases}$$

peut être imposé comme contrainte en plus de (2), ce qui donne les poids

$$w_i = \frac{\hat{N}}{N} d_i + \left( \mathbf{X} - \frac{\hat{N}}{N} \mathbf{X}^p \right)'$$

$$\left\{ \sum_{j \in A} d_j' (\mathbf{x}_j - \mathbf{X}^p) (\mathbf{x}_j - \mathbf{X}^p)' \right\}^{-1} d_i' (\mathbf{x}_i - \mathbf{X}^p), \quad (5)$$

$$\hat{Y}^p = \sum_{i=1}^n d_i' y_i, \quad (1)$$

L'estimateur de Horvitz-Thompson (HT) de la forme

où  $d_i = 1/\pi_i$  est le poids d'échantillonnage et  $\pi_i$  est la probabilité d'inclusion de premier ordre, est sans biais pour  $\bar{Y}$ . Toutefois, il n'utilise pas l'information fournie par  $\mathbf{X}$ . Selon Kott (2006), un estimateur par calage peut être défini comme l'estimateur de la forme

$$\hat{Y}^w = \sum_{i \in A} w_i y_i$$

où les poids  $w_i$  satisfont

$$\sum_{i \in A} w_i \mathbf{x}_i = \mathbf{X} \quad (2)$$

et  $\hat{Y}^w$  est asymptotiquement sans biais par rapport au plan. L'estimation par calage est aujourd'hui très répandue dans les enquêtes par sondage, parce qu'elle assure la cohérence des résultats entre diverses enquêtes et améliore souvent l'efficacité (Särndal 2007).

- Rizzo, L., Kalton, G. et Brick, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.
- Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York : John Wiley & Sons, Inc.
- Särndal, C.-E., et Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 4, 251-260.
- Schouten, B. (2007). A selection strategy for weighting variables under a not-missing-at-random assumption. *Journal of Official Statistics*, 23, 51-68.
- Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 107-121.
- Thomsen, I., Kleven, Ø., Wang, J.H. et Zhang, L.C. (2006). Coping with decreasing response rates in Statistics Norway. Recommended practice for reducing the effect of nonresponse. Rapport 2006/29. Oslo : Statistics Norway.

11. Conclusion

Dans le présent article, nous examinons des situations d'enquête dans lesquelles de nombreux vecteurs auxiliaires (vecteurs  $\mathbf{x}$ ) possibles peuvent être créés et considérons l'utilisation de l'estimateur par calage  $\hat{X}_{CAL}$ . Pour tout vecteur  $\mathbf{x}$  donné, un certain biais inconnu persiste dans  $\hat{X}_{CAL}$ ; nous souhaitons, par un choix approprié du vecteur  $\mathbf{x}$ , rendre le biais aussi faible que possible. Donc, nous examinons le ratio des biais défini par (4.2) et (4.3). Nous exprimons, dans (5.8) à (5.10), la composante  $\Delta_q$  du ratio des biais sous la forme d'un produit de mesures statistiques faciles à interpréter. Cela nous mène à proposer plusieurs variantes d'indicateurs de biais pouvant être utilisées pour évaluer les divers vecteurs  $\mathbf{x}$  en ce qui a trait à leur capacité de réduire efficacement le biais. Nous étudions en particulier l'indicateur  $H_1$  donné par (5.12). Il fonctionne très bien, mais est axé sur une variable étudiée  $y$  particulière. Toutefois, une enquête gouvernementale type comprend un grand nombre de variables étudiées et, pour des raisons pratiques, il est souhaitable d'utiliser le même vecteur  $\mathbf{x}$  pour estimer tous les totaux  $y$ . Un compromis devient donc nécessaire. Nous soutenons que l'indicateur  $H_3$  donné par (5.12), répond à ce besoin; il dépend des valeurs  $\mathbf{x}_k$ , mais ne dépend d'aucune donnée  $y$ . L'élaboration d'autres indicateurs (que  $H_3$ ) pour la « situation comportant de nombreuses variables  $y$  » sera le sujet de futurs travaux de recherche. L'examen, pour la sélection pas à pas des variables  $\mathbf{x}$  avec l'indicateur  $H_1$ , d'autres algorithmes que celui utilisé à la section 9 sera aussi le sujet de travaux de recherche à venir.

Remerciements

Les auteurs remercient les examinateurs et le rédacteur associé de leurs commentaires qui ont contribué à l'amélioration du présent article.

Bibliographie

Deville, J.-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des Journées de Méthodologie*, I.N.S.E.E., Paris.

Eittinge, J., et Vansaneh, I. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23, 37-45.

Kalton, G., et Flores-Cervantes, I. (2003). Weighing methods. *Journal of Official Statistics*, 19, 81-98.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 149-160.

5 000 résultats ( $s, r$ ) ont été réalisés, d'abord avec un échantillon de taille  $n = 1\,000$ , puis avec un échantillon de taille  $n = 2\,000$  (l'ensemble de réponses  $r$  est réalisé conformément à l'une des quatre distributions des réponses, en déclarant l'unité  $k$  « répondante » à la suite d'un essai de Bernoulli avec la probabilité spécifiée  $\theta_k$ ). Nous avons calculé le taux de succès comme étant la proportion de l'ensemble des résultats ( $s, r$ ) dans laquelle l'indication correcte se concrétise dans une confrontation de deux vecteurs  $\mathbf{x}$  différents. Nous avons procédé à plusieurs comparaisons par paire de cette sorte. Des résultats types sont présentés au tableau 10.4 pour  $\text{InExp}(10 + x_1 + x_2)$ . L'entrée supérieure dans une cellule du tableau montre le taux de succès en % pour  $n = 1\,000$ , et l'entrée inférieure, le taux pour  $n = 2\,000$ . La valeur de *biais rel* pour les vecteurs en question est entre parenthèses.

Nous préférons les « tests sévères », c'est à dire les confrontations de vecteurs pour lesquels la différence absolue de *biais rel* est faible, parce que la décision correcte est alors plus difficile à obtenir. Il n'existe a priori aucune raison qu'un des indicateurs donne systématiquement de meilleurs résultats que les autres dans la présente étude. Dans les cinq tests sévères du tableau 10.4,  $H_1$  produit, dans l'ensemble, un meilleur taux de succès que  $H_2$  et  $H_3$ . Le taux de succès de  $H_1$  s'améliore si l'on double la taille d'échantillon et a tendance, comme prévu, à être plus élevé quand les valeurs de *biais rel* sont plus écartées. Le cas  $4G + 8G$  c.  $8G + 8G$  compare les vecteurs  $\mathbf{x}$  emboîtés, de sorte que l'on sait avant l'expérience que  $H_2$  et  $H_3$  donnent des taux de succès parfaits.

Tableau 10.4

Certains comparaisons par paire des vecteurs auxiliaires; pourcentage de résultats avec indication correcte, pour les indicateurs  $H_1, H_2$  et  $H_3$ . Entre parenthèses, *biais rel* en %. Entrée supérieure:  $n = 1\,000$ , entrée inférieure:  $n = 2\,000$ . Distribution des réponses  $\text{InExp}(10 + x_1 + x_2)$

Cellules comparées	Pourcentage de résultats avec indication correcte		
	$H_1$	$H_2$	$H_3$
$4G + 8G(0,5)$ c.	90,0	100,0	100,0
$8G + 8G(0,2)$	96,4	100,0	100,0
$4G + 2G(1,8)$ c.	66,8	86,0	70,7
$2G + 8G(1,5)$	74,2	89,0	67,4
$1G + 8G(4,1)$ c.	74,3	70,3	45,0
$8G + 1G(3,4)$	82,8	78,0	43,3
$4G + 1G(4,1)$ c.	90,6	61,4	83,9
$2G + 2G(3,2)$	97,0	68,8	92,3
$1G + 2G(7,3)$ c.	77,4	77,4	34,5
$2G + 1G(6,5)$	85,9	85,9	28,8



Le tableau 10.2 pour  $\text{InExp}(10 + x_1 + x_2)$  et le tableau 10.3 pour  $\text{InExp}(10 + y)$  montrent comment  $\text{Moy}(H_1)$ ,  $\text{Moy}(H_2)$  et  $\text{Moy}(H_3)$  classent les 16 vecteurs  $x$ , représentés par leur valeur de *biais rel*. Pour mesurer le succès du classement, nous avons calculé le coefficient de corrélation de rangs de Spearman, désigné par *corrang*, entre *biais rel* et la valeur de l'indicateur, basé sur les 16 valeurs de tableaux donne  $|\text{corrang}|=1$ , pour le classement parfait. Pour les données utilisées,  $|\text{corrang}|$  est aussi presque égal à un pour  $\text{Moy}(H_2)$  et  $\text{Moy}(H_3)$  (plus généralement, le classement obtenu avec  $H_2$  et  $H_3$  peut être bon, mais dépend des données).

Tableau 10.2

Valeur, par ordre croissant, de <i>biais rel</i> en %, et valeur et rang correspondants de $\text{Moy}(H_1) \times 10^3$ , $\text{Moy}(H_2) \times 10^3$ et $\text{Moy}(H_3) \times 10^3$			
Distribution des réponses $\text{InExp}(10 + x_1 + x_2)$			
<i>biais rel</i>	$\text{Moy}(H_1) \times 10^3$	$\text{Moy}(H_2) \times 10^3$	$\text{Moy}(H_3) \times 10^3$
0,2	101	127	232
0,5	99	119	225
0,5	98	118	224
0,8	96	109	217
1,3	93	109	216
1,5	91	105	213
1,8	89	98	207
1,9	88	94	205
3,2	78	80	192
3,4	76	90	188
4,1	70	84	190
4,1	70	84	190
5,0	64	70	179
5,0	64	70	179
6,4	52	52	146
7,3	46	46	156
13,2	0	0	0
<i>Corrang</i>	(16)	(16)	(16)
	(1)	(1)	(1)
	(2)	(2)	(2)
	(3)	(3)	(3)
	(4)	(4)	(4)
	(5)	(5)	(5)
	(6)	(6)	(6)
	(7)	(7)	(7)
	(8)	(8)	(8)
	(9)	(9)	(9)
	(10)	(10)	(10)
	(11)	(11)	(11)
	(12)	(12)	(12)
	(13)	(13)	(13)
	(14)	(14)	(14)
	(15)	(15)	(15)
	(16)	(16)	(16)

Il existe un contraste appréciable entre les résultats pour *biais rel* pour les deux distributions des réponses dans les tableaux 10.2 et 10.3. Le meilleur des vecteurs auxiliaires laisse un biais considérablement plus important pour la distribution  $\text{InExp}(10 + y)$  non ignorable que pour la distribution  $\text{InExp}(10 + x_1 + x_2)$ . Cela n'est pas étonnant et il est important de noter qu'une réduction importante du biais est obtenue pour le cas non ignorable également. Dans la simulation, le surajustement mentionné à la section 4,  $\Delta_A > \Delta_T > 0$  (quand  $(\hat{Y}^{\text{Exp}})$  présente un biais positif) ou  $\Delta_A < \Delta_T < 0$  (quand  $\hat{Y}^{\text{Exp}}$  présente un biais négatif), se produit pour certains résultats ( $s, r$ ). La fréquence varie selon la force du vecteur auxiliaire et diffère pour les diverses distributions des réponses. La cellule pour laquelle ce surajustement a le plus de chance de se produire est  $8G + 8G$ , qui comprend le plus puissant des 16 vecteurs auxiliaires. Pour  $\text{InExp}(10 + x_1 + x_2)$ , le biais est presque

Une question qui n'est pas traitée dans les tableaux 10.2 et 10.3 est celle de savoir combien de fois, sur une longue série de résultats ( $s, r$ ), un indicateur donne  $H(x_k)$  réussit à désigner correctement le vecteur  $x$  préféré. Pour répondre à cette question, comparons les deux vecteurs  $x_{1k}$  et  $x_{2k}$ . Si la valeur absolue du biais de  $\hat{Y}^{\text{CAL}}(x_{2k})$  est plus petite que celle du biais de  $\hat{Y}^{\text{CAL}}(x_{1k})$ , nous aimerions observer que l'inégalité  $H(x_{2k}) \geq H(x_{1k})$  est vérifiée pour une grande majorité des résultats ( $s, r$ ), parce qu'alors, l'indicateur  $H(\cdot)$  produira avec une probabilité élevée la décision correcte de préférer  $x_{2k}$ . Comme  $H(x_k)$  présente une variabilité d'échantillonnage, son taux de succès (le taux d'indication correcte) dépend de la taille de l'échantillon et nous nous attendons à ce qu'il augmente avec cette taille.

Tableau 10.3  
Valeur, par ordre croissant, de *biais rel* en %, et valeur et rang correspondants de  $\text{Moy}(H_1) \times 10^3$ ,  $\text{Moy}(H_2) \times 10^3$  et  $\text{Moy}(H_3) \times 10^3$ , pour 16 vecteurs auxiliaires. Ligne inférieure : valeur des corrélations de rangs de Spearman, *corrang*. Distribution des réponses  $\text{InExp}(10 + y)$

Valeur, par ordre croissant, de <i>biais rel</i> en %, et valeur et rang correspondants de $\text{Moy}(H_1) \times 10^3$ , $\text{Moy}(H_2) \times 10^3$ et $\text{Moy}(H_3) \times 10^3$			
Distribution des réponses $\text{InExp}(10 + y)$			
<i>biais rel</i>	$\text{Moy}(H_1) \times 10^3$	$\text{Moy}(H_2) \times 10^3$	$\text{Moy}(H_3) \times 10^3$
3,6	74	74	165
3,9	71	71	158
4,0	71	71	156
4,3	68	68	149
4,4	68	68	153
4,9	64	64	142
4,9	64	64	146
4,9	63	63	147
5,3	60	60	143
5,4	60	60	137
6,0	55	55	132
6,2	53	53	128
7,2	46	46	122
7,9	41	41	111
7,9	41	41	111
7,9	40	40	109
9,6	27	27	90
13,1	0	0	0
<i>Corrang</i>	(16)	(16)	(16)
	(1)	(1)	(1)
	(2)	(2)	(2)
	(3)	(3)	(3)
	(4)	(4)	(4)
	(5)	(5)	(5)
	(6)	(6)	(6)
	(7)	(7)	(7)
	(8)	(8)	(8)
	(9)	(9)	(9)
	(10)	(10)	(10)
	(11)	(11)	(11)
	(12)	(12)	(12)
	(13)	(13)	(13)
	(14)	(14)	(14)
	(15)	(15)	(15)
	(16)	(16)	(16)

Nous apportons certain éclaircissements au sujet de cette question en prolongeant l'expérience de Monte Carlo :

L'indicateur  $H_0$  classera les  $4 \times 4 = 16$  vecteurs auxiliaires correctement pour toute distribution des réponses (les probabilités de réponse n'étant pas toutes constantes, comme il est mentionné plus bas). Le tableau 10.1 illustre (8.4), en fonction de  $H_1 = |H_0|$ : la variation, de n'importe quelle cellule à n'importe quelle autre, de la valeur de  $\text{Moy}(H_1)$  (l'estimation Monte Carlo) de la valeur prévue de  $(H_1)$  est accompagnée d'une variation proportionnelle de la valeur de *biais rel*. La même proportionnalité a été observée pour les trois autres distributions des réponses. Nous aurons pu choisir d'autres distributions des réponses pour illustrer la même propriété.

Tableau 10.1  
*Biais rel* en % et, entre parenthèses, la valeur de  $\text{Moy}(H_1) \times 10^3$  pour 16 vecteurs auxiliaires  $x_k$ . Distribution des réponses  $\text{IncExp}(10 + x_1 + x_2)$

Groupes basés sur $x_k$	Groupes basés sur $x_2 k$			
	1G	2G	4G	8G
1G	0,2 (101)	0,5 (99)	1,3 (93)	3,4 (76)
2G	0,5 (98)	0,9 (96)	1,8 (89)	4,1 (70)
4G	1,5 (91)	1,9 (88)	3,2 (78)	6,5 (52)
8G	4,1 (70)	5,0 (64)	7,3 (46)	13,2 (0)

La distribution des réponses avec une probabilité de réponse constante  $\theta_k$  pour tout  $k$  est un cas particulier. L'estimateur par calage  $\hat{Y}^{\text{CAL}}$  basé sur tout vecteur  $x_k$  présente alors un biais nul (quasiment) et cela inclut l'estimateur élémentaire  $\hat{Y}^{\text{EXP}}$  avec  $x_k = 1$ . Le résultat 8.3 continue d'être valide, indiquant dans ce cas que  $E^{pg}(H_0) \approx \text{biais}(\hat{Y}^{\text{CAL}}) \approx \text{biais}(\hat{Y}^{\text{EXP}}) \approx 0$ . Dans le contexte de la simulation de la présente section, si  $\theta_k = 0,70$  pour tout  $k$  est considérée comme une distribution supplémentaire des réponses, le tableau 10.1 contenant chacune des 16 cellules montre des valeurs presque nulles de *biais rel* en % et de  $\text{Moy}(H_1) \times 10^3$ , de la cellule la plus faible (1G + 1G) jusqu'à la cellule du vecteur  $x$  le plus puissant (8G + 8G). Il n'existe aucun biais devant être éliminé par une amélioration du vecteur  $x$ . Si, en pratique, l'indicateur ( $H_1$ ) ne réagit pas à un agrandissement du vecteur  $x$ , il n'y a aucune raison de pousser la recherche au delà de la formule vectorielle la plus simple. Cette situation peut s'interpréter de trois façons : la variable  $y$  en question ne comporte aucun biais de non réponse, ou la probabilité de réponse est presque constante, ou aucun des vecteurs  $x$  disponibles n'est capable de réduire un biais existant.

Par souci de concision, nous ne montrons pas les tableaux correspondants pour  $\text{Moy}(H_2)$  et  $\text{Moy}(H_3)$ . Par nécessité mathématique, les deux quantités augmentent dans les transitions emboîtées. Nous ne présentons pas non plus les analogues du tableau 10.1 pour les trois autres distributions des réponses, car les profils sont comparables.

les probabilités de réponse : croissantes (IncExp) par opposition à décroissantes (DecExp), dépendantes des valeurs de  $x$  uniquement par opposition à dépendantes des valeurs de  $y$  uniquement. Dans la deuxième et la quatrième options, la réponse dépend directement de la variable  $y$  et pourrait donc être appelée « purement non ignorable ».

Nous avons généré  $J = 5\,000$  résultats  $(s, r)$ , où  $s$  de taille  $n = 1\,000$  est tiré de  $N = 6\,000$  par échantillonnage aléatoire simple et, pour chaque  $s$  donné, l'ensemble de réponses  $r$  est réalisé par chacune des quatre distributions des réponses. Autrement dit, pour  $k \in s$ , nous avons effectué un essai de Bernoulli avec la probabilité spécifiée  $\theta_k$  d'inclusion dans l'ensemble des réponses  $r$ . Les essais de Bernoulli sont indépendants.

Pour chaque distribution des réponses, pour chacun des 16 vecteurs  $x$ , et pour chaque résultat  $(s, r)$ , nous avons calculé l'écart relatif  $\text{ER} = (\hat{Y}^{\text{CAL}} - Y)/Y$ , où  $\hat{Y}^{\text{CAL}}$  est donné par (2.4) et  $Y = \sum_U y_k$  est le total de  $y$  cible, connu dans les conditions de cette expérience (alternativement, nous avons utilisé  $\hat{Y}^{\text{CAL}}$  donné par (2.5), mais, comme prévu, la différence de biais comparativement à  $\hat{Y}^{\text{CAL}}$  était négligeable). Nous avons également calculé les indicateurs  $H_i, i = 0, 1, 2, 3$ , donnés par (5.11) et (5.12). Nous avons calculé les mesures sommatives suivantes :

$$\text{biais rel} = \text{Moy}(\text{ER}) = \frac{1}{J} \sum_{j=1}^J \text{ER}_j,$$

$$\text{Moy}(H_i) = \frac{1}{J} \sum_{j=1}^J H_{ij} \quad \text{pour } i = 0, 1, 2, 3$$

où  $j$  indique la valeur calculée pour le  $j^{\text{e}}$  résultat,  $j = 1, 2, \dots, 5\,000 = J$ . Pour chaque distribution des réponses, nous avons donc obtenu la valeur *biais rel* (qui est la mesure Monte Carlo du biais relatif  $(\hat{Y}^{\text{CAL}} - Y)/Y$ ) et 16 valeurs de  $\text{Moy}(H_i)$  (qui est la mesure Monte Carlo de  $E^{pd}(H_i)$ ),  $i = 0, 1, 2, 3$ , où  $p$  désigne l'échantillonnage aléatoire simple et  $q$  représente l'une des quatre distributions des réponses.

Le tableau 10.1 montre, pour  $\text{IncExp}(10 + x_1 + x_2)$ , *biais rel* en % et  $\text{Moy}(H_1) \times 10^3$  pour les 16 vecteurs  $x$ . Pour la cellule 1G + 1G, avec le vecteur  $x_k = 1$ , les quatre quantités moyennes ( $\text{Moy}$ ) sont nulles et *biais rel* atteint son niveau le plus élevé, soit 13,2%. À l'autre extrême, la cellule 8G + 8G représente le niveau le plus élevé d'information : elle produit la valeur la plus élevée pour  $\text{Moy}(H_1)$ , et *biais rel* atteint sa valeur la plus faible, soit 0,2% ; presque tout le biais est éliminé (à part une différence de signe éventuelle,  $\text{Moy}(H_0)$  et  $\text{Moy}(H_1)$ ) étaient égaux pour toutes les cellules).  
Le résultat (8.4), qui est vérifié pour toute distribution des réponses et tout plan d'échantillonnage, indique que



nouveau groupe de 1 500, et ainsi de suite ; la valeur vectorielle associée à l'unité  $k$  est  $\gamma_{(x_1;4)k}$ . Dans le mode 2G, l'unité  $k$  possède la valeur vectorielle  $\gamma_{(x_1;2)k} = (1, 0)'$  pour les 3 000 unités possédant les plus grandes valeurs  $x_1$  et  $\gamma_{(x_1;2)k} = (0, 1)'$  pour les autres. Dans le dernier mode 1G, les 6 000 unités sont toutes regroupées, toute l'information contenue dans  $x_1$  est abandonnée, et  $\gamma_{(x_1;1)k} = 1$  pour tout  $k$ . Les 6 000 valeurs  $x_{2k}$  ont été regroupées selon la même procédure en les quatre modes 8G, 4G, 2G et 1G. L'appartenance correspondante de l'unité  $k$  à ces groupes est codée par les vecteurs  $\gamma_{(x_2;8)k}, \gamma_{(x_2;4)k}, \gamma_{(x_2;2)k}$  et  $\gamma_{(x_2;1)k}$ . Les  $4 \times 4 = 16$  vecteurs auxiliaires  $\mathbf{x}_k$  différents tiennent compte des deux sortes d'information de groupe ; les deux vecteurs  $\gamma$  sont placés côte à côte (par opposition à croisés), le résultat étant un calage sur deux marges, comme indiqué par le signe « + ». Donc, pour le cas désigné 8G + 8G, l'unité  $k$  possède la valeur vectorielle auxiliaire  $\mathbf{x}_k = (\gamma'_{(x_1;8)k}, \gamma'_{(x_2;8)k})'(-1)$  ou  $(-1)$  indique qu'une catégorie est exclue dans  $\gamma_{(x_1;8)k}$  ou dans  $\gamma_{(x_2;8)k}$  pour éviter une matrice singulière dans les calculs, ce qui donne à  $\mathbf{x}_k$  la dimension  $8 + 8 - 1 = 15$ . Le cas 8G + 8G est celui dont le contenu informationnel est le plus important. À l'autre extrême, le cas 1G + 1G exclut toute l'information contenue dans  $x$  et  $\mathbf{x}_k = 1$  pour tout  $k$ . Il existe 14 cas intermédiaires de contenu informationnel. Par exemple, 4G + 2G possède le vecteur  $\mathbf{x}_k = (\gamma'_{(x_1;4)k}, \gamma'_{(x_2;2)k})'(-1)$  de dimension  $4 + 2 - 1 = 5$  ; 4G + 1G possède  $\mathbf{x}_k = (\gamma'_{(x_1;4)k}, 1)'(-1)$  de dimension 4 (il existe une interaction non négligeable entre  $x_1$  et  $x_2$  dans cette expérience, mais nous limitons cette dernière aux vecteurs  $\mathbf{x}$  sans interaction, ce qui évite le risque d'obtenir des groupes dans lesquels la fréquence est faible).

Nous discutons ici des résultats pour quatre distributions des réponses. Les probabilités de réponse  $\theta_k, k = 1, 2, \dots, N = 6\,000$  de ces distributions ont été spécifiées comme il suit :

$$\begin{aligned} \text{IncExp}(10 + x_1 + x_2), & \text{ avec } \theta_k = 1 - e^{-c(10 + x_{1k} + x_{2k})}, & \text{ où } c = 0,04599 \\ \text{IncExp}(10 + y), & \text{ avec } \theta_k = 1 - e^{-c(10 + y_k)}, & \text{ où } c = 0,06217 \\ \text{DecExp}(x_1 + x_2), & \text{ avec } \theta_k = e^{-c(x_{1k} + x_{2k})}, & \text{ où } c = 0,01937 \\ \text{DecExp}(y), & \text{ avec } \theta_k = e^{-cy_k}, & \text{ où } c = 0,03534. \end{aligned}$$

La constante  $c$  était ajustée dans les quatre cas de manière à obtenir une probabilité de réponse moyenne de  $\bar{\theta}_U = \sum \theta_k / N = 0,70$ . Dans les deux premiers, la valeur 10 (plutôt que 0) a été utilisée pour éviter une fréquence élevée de faibles probabilités de réponse  $\theta_k$ . Ces quatre options représentaient des caractéristiques contrastantes pour

deux variables auxiliaires continues,  $x_1$  et  $x_2$ . Les valeurs  $(y_k, x_{1k}, x_{2k})$  pour  $k = 1, 2, \dots, 6\,000$  ont été créées en trois étapes, comme il suit. Étape 1 (variable  $x_1$ ) : les 6 000 valeurs  $x_{1k}$  ont été obtenues comme des résultants indépendants de la variable aléatoire de loi gamma  $\Gamma(a, b)$  en donnant aux paramètres les valeurs  $a = 2, b = 5$ . La moyenne et la variance des 6 000 valeurs réalisées  $x_{1k}$  étaient de 10,0 et 49,9, respectivement. Étape 2 (variable  $x_2$ ) : pour l'unité  $k$ , avec la valeur  $x_{1k}$  fixée à l'étape 1, une valeur  $x_{2k}$  est réalisée en tant que résultat de la variable aléatoire gamma dont les paramètres sont tels que l'espérance et la variance conditionnelles de  $x_{2k}$  sont  $\alpha + \beta x_{1k} + K h(x_{1k})$  et  $\sigma^2 x_{1k}$ , respectivement, où  $h(x_{1k}) = x_{1k}(x_{1k} - \mu_{x_1}) - 3\mu_{x_1}$  avec  $\mu_{x_1} = 10$ . Nous avons utilisé les valeurs  $\alpha = 1, \beta = 1, k = 1, \sigma^2 = 25$ . Le terme polynomial  $K h(x_{1k})$  donne une forme légèrement non linéaire au tracé de  $(x_{2k}, x_{1k})$ , pour éviter une relation exactement linéaire. La moyenne et la variance des 6 000 valeurs réalisées  $x_{2k}$  étaient égales à 11,0 et 210,0, respectivement. Le coefficient de corrélation entre  $x_1$  et  $x_2$ , calculé sur les 6 000 couples  $(x_{1k}, x_{2k})$ , était de 0,48. Étape 3 (variable étudiée  $y$ ) : pour l'unité  $k$ , avec les valeurs  $x_{1k}$  et  $x_{2k}$  fixées aux étapes 1 et 2, une valeur  $y_k$  est réalisée en tant que résultat de la variable aléatoire gamma dont les paramètres sont tels que l'espérance et la variance conditionnelles de  $y_k$  sont  $c_0 + c_1 x_{1k} + c_2 x_{2k}$  et  $\sigma_0^2 (c_1 x_{1k} + c_2 x_{2k})$ , respectivement. Nous avons utilisé  $c_0 = 1, c_1 = 0,7, c_2 = 0,3$  et  $\sigma_0^2 = 2$ . La moyenne et la variance des 6 000 valeurs réalisées  $y_k$  étaient égales à 11,4 et 86,5, respectivement. Le coefficient de corrélation entre  $y$  et  $x_1$ , calculé sur les 6 000 couples  $(y_k, x_{1k})$ , était de 0,76 ; celui entre  $y$  et  $x_2$ , calculé sur les 6 000 couples  $(y_k, x_{2k})$ , était de 0,73.

Chaque des deux variables  $x$  a été transformée ensuite en quatre variantes de mode de groupement, désignées 8G, 4G, 2G et 1G, produisant  $4 \times 4 = 16$  vecteurs auxiliaires  $\mathbf{x}_k$  différents. Nous avons classé les 6 000 valeurs  $x_{1k}$  de la variable  $x_1$  par ordre de grandeur et avons formé huit groupes de taille égale. Le premier groupe comprenait les 750 unités ayant les valeurs les plus grandes de  $x_{1k}$ , le deuxième, les 750 unités suivantes du classement par ordre de grandeur, et ainsi de suite, jusqu'au huitième groupe. Dans ce mode de groupement 8G de  $x_1$ , à l'unité  $k$  est affectée la valeur vectorielle  $\gamma_{(x_1;8)k}$ , de dimension huit avec sept entrées « 0 » et une seule entrée « 1 » pour coder l'appartenance de  $k$  au groupe. Ensuite, des fusions successives de groupes sont effectuées, de sorte que deux groupes adjacents définissent toujours un nouveau groupe, chaque fois en doublant la taille du groupe. Donc, pour le mode 4G, la fusion des groupes 1 et 2 place les unités ayant les 1 500 plus grandes valeurs  $x_{1k}$  dans un premier nouveau groupe ; la fusion des groupes 3 et 4 produit le deuxième



L'ensemble des six premières variables sélectionnées avec  $H_3$  contient trois des mêmes variables que l'ensemble correspondant de six variables sélectionnées avec  $H_1$ . Les deux profils de sélection assez différents ne sont pas en contradiction, parce que  $H_1$  est axé spécifiquement sur la variable  $y_{Occupé(e)}$ , tandis que  $H_3$  est un indicateur de compromis, indépendant de toute variable  $y$ . Faute d'espace, nous ne présentons pas les résultats de la sélection de cet indicateur  $H_2$ . Le profil de sélection de cet indicateur ressemble davantage à celui de  $H_3$  qu'à celui de  $H_1$ . Parmi les six premières variables sélectionnées avec  $H_2$ , quatre sont parmi les six premières sélectionnées avec  $H_3$ . À titre de commentaire général, nous pensons que dans de nombreuses situations pratiques, l'utilisation de plus de six variables est inutile et que la sélection des quelques premières devient très importante.

### 10. Validation empirique par simulation pour une population synthétique

La théorie présentée aux sections précédentes ne comporte aucune hypothèse quant à la distribution des réponses, qui est inconnue. Le plan d'échantillonnage est arbitraire et les probabilités d'inclusion connues sont prises en compte. Pour l'expérience décrite à la présente section, nous spécifions plusieurs distributions des réponses pour lesquelles nous précisons une valeur positive de la probabilité de réponse  $\theta_k$  pour chaque  $k \in L$ . Autrement dit, pour la probabilité spécifiée  $\theta_k$ , la valeur  $y_k$  est enregistrée dans l'expérience et pour la probabilité  $1 - \theta_k$ , cette valeur est manquante. Nous constatons que les indicateurs  $H_0$  (ou  $H_1 = |H_0|$ ) définis en (5.11) classent les divers vecteurs  $x$  dans l'ordre correct de préférence pour toutes les distributions des réponses prises en considération, conformément aux résultats théoriques (8.3) et (8.4). Nous confirmons que, sur une longue série de résultats  $(s, r)$ , la moyenne de  $H_0 = \Delta^A / S_y = -R_{y,m} \times cv_m$  suit le biais des estimateurs par calage, mesuré par la moyenne de  $\bar{Y}_{CAL} - Y$ , de manière essentiellement parfaitement linéaire, quand le vecteur  $x$  correspond successivement aux 16 formules différentes. Nous examinons aussi les indicateurs  $H_2$  et  $H_3$  définis en (5.12) et constatons dans cette expérience qu'ils sont fortement reliés au biais de  $\bar{Y}_{CAL}$ .

Nous avons expérimenté plusieurs populations synthétiques et obtenu des conclusions similaires. Nous présentons ici les résultats pour une population synthétique de taille  $N = 6\,000$ , avec les valeurs créées  $(y_k, x_k, \theta_k)$  pour  $k = 1, 2, \dots, N = 6\,000$ , pour 16 formules catégoriques différentes de  $x_k$ , et quatre façons distinctes d'attribuer la probabilité  $\theta_k$ .

Les 16 vecteurs  $x$  de variables catégoriques ont été obtenus en regroupant les valeurs générées  $x_{1k}$  et  $x_{2k}$  de

$\bar{Y}_{CAL}$  et ERC commencent à augmenter. À l'étape 6, ERC atteint sa valeur la plus faible, soit 0,5, puis commence à augmenter, illustrant le fait que l'ajout de toutes les variables  $x$  disponibles pourrait ne pas représenter la meilleure approche. Le point de virage de  $H_1$  et le point auquel ERC est le plus proche de zéro coïncident dans le présent exemple, mais il n'est généralement pas ainsi. En outre, dans des conditions réelles d'enquête, ERC est inconnu, de même que l'étape à laquelle il est le plus proche de zéro.

Le tableau 9.2 donne la sélection pas à pas avec l'indicateur  $H_3$ . La valeur de ce dernier augmente à chaque étape, mais à une vitesse qui finit par plafonner et les variations successives de  $\bar{Y}_{CAL}$  deviennent négligeables. Ces résultats donnent à penser qu'il faut s'arrêter après six étapes, au moment où ERC = 2,8. Dans aucune des 12 étapes ERC ne s'approche autant de zéro que la valeur de 0,5 obtenue avec  $H_1$  après six étapes. À cet égard,  $H_1$  est meilleur que  $H_3$ , dans le présent exemple. Quand les 12 variables  $x$  sont toutes sélectionnées, ERC atteint la valeur finale de 2,6 dans les deux tableaux.

Tableau 9.2

Sélection pas à pas ascendante, indicateur  $H_1$ , variable dichotomique *Occupé(e)*. Valeurs successives de  $H_1 \times 10^3$ , de  $\bar{Y}_{CAL}$  en milliers et de ERC =  $(\bar{Y}_{CAL} - \bar{Y}_{FUL}) / \bar{Y}_{FUL} \times 10^2$ . Pour comparaison,  $\bar{Y}_{EXP} \times 10^{-3} = 4\,719$ ;  $\bar{Y}_{FUL} \times 10^{-3} = 4\,265$

Variable auxiliaire ajoutée	$H_1 \times 10^3$	$\bar{Y}_{CAL} \times 10^3$	ERC
Catégorie de revenu (3)	76	4 458	4,5
Niveau de scolarité (3)	107	4 350	2,0
Présence d'enfants (2)	114	4 326	1,4
Logement urbain (2)	118	4 310	1,1
Sexe (2)	123	4 296	0,7
Etat matrimonial (2)	125	4 286	0,5
Jours de chômage (3)	121	4 301	0,9
Mois de prestations de maladie (3)	120	4 305	1,0
Niveau d'endettement (3)	115	4 322	1,3
Grappe de codes postaux (6)	109	4 343	1,8
Pays de naissance (2)	103	4 363	2,3
Groupe d'âge (4)	99	4 377	2,6

Sélection pas à pas ascendante, indicateur  $H_3$ , variable dichotomique *Occupé(e)*. Valeurs successives de  $H_3 \times 10^3$ , de  $\bar{Y}_{CAL}$  en milliers, de ERC =  $(\bar{Y}_{CAL} - \bar{Y}_{FUL}) / \bar{Y}_{FUL} \times 10^2$ . Pour comparaison,  $\bar{Y}_{EXP} \times 10^{-3} = 4\,719$ ;  $\bar{Y}_{FUL} \times 10^{-3} = 4\,265$

Variable auxiliaire ajoutée	$H_3 \times 10^3$	$\bar{Y}_{CAL} \times 10^3$	ERC
Niveau de scolarité (3)	186	4 520	6,0
Grappe de codes postaux (6)	250	4 505	5,6
Pays de naissance (2)	281	4 498	5,5
Catégorie de revenu (3)	298	4 369	2,4
Groupe d'âge (4)	354	4 399	3,1
Sexe (2)	364	4 384	2,8
Logement urbain (2)	374	4 378	2,6
Niveau d'endettement (3)	381	4 364	2,3
Mois de prestations de maladie (3)	384	4 380	2,7
Présence d'enfants (2)	387	4 379	2,7
Etat matrimonial (2)	388	4 379	2,7
Jours de chômage (3)	388	4 377	2,6

pour cette méthode, qui est, à cet égard, semblable à la nôtre.

### 9. Choix des variables auxiliaires pour l'enquête pilote suédoise sur les jeux de hasard et le jeu compulsif

Nous avons utilisé un ensemble de données d'enquête réelles pour illustrer l'utilisation des indicateurs  $H_1$ ,  $H_2$  et  $H_3$  en vue de construire le vecteur  $x$ . En 2008, l'Institut national de santé publique de la Suède (*Svenska Folkhälsoinstitutet*) a réalisé une enquête pilote pour étudier la portée de la participation à des jeux de hasard et les caractéristiques des joueurs compulsifs. L'échantillonnage et le calage des poids ont été effectués par Statistics Sweden. Nous illustrons l'utilisation des indicateurs dans cette enquête, pour laquelle un échantillon aléatoire simple stratifié  $s$  de  $n = 2\,000$  personnes a été tiré du registre de la population totale (RPT) de la Suède. Les strates ont été définies par recoupement de la région de résidence et du groupe d'âge. Chacune des six régions a été définie comme une grappe de zones postales considérées comme étant semblables en ce qui a trait à des variables telles que le niveau de scolarité, le pouvoir d'achat, le type de logement, et l'origine étrangère. Les quatre groupes d'âge utilisés étaient 16 à 24 ans, 25 à 34 ans, 35 à 64 ans et 65 à 84 ans.

Le taux de réponse global pondéré était de 50,8 %. La non réponse, plus ou moins prononcée selon le domaine d'intérêt, interfère avec l'objectif de précision établi. Un important réservoir de variables auxiliaires possibles était disponible pour l'enquête, y compris des variables du RPT, des variables du registre de l'éducation et un sous ensemble de celles comprises dans une autre grande base de données de Statistics Sweden, appelée LISA. Pour le présent exemple, nous avons créé un fichier de données dans lequel nous avons sélectionné 13 variables catégoriques. Nous avons utilisé 12 d'entre elles comme des variables  $x$  et la treizième, la variable dichotomique *Occupée(e)*, comme variable étudiée. Les valeurs de toutes les variables sont disponibles pour toutes les unités  $k \in s$ . La réponse ( $k \in r$ ) ou la non réponse ( $k \in s - r$ ) à l'enquête est également indiquée dans le fichier de données.

Les variables de nature continue sont utilisées comme des variables groupées, de sorte que les 12 variables  $x$  sont catégoriques et de type  $x_k^0$ , comme il est défini à la section 2 (comme la plupart des variables sont disponibles pour la totalité de la population, elles pourraient être de type  $x_k^*$ , mais puisque l'effet du biais a peu d'importance, nous les avons utilisées comme des variables  $x_k^0$ ). La valeur de la variable étudiée,  $y_k = 1$  si l'unité  $k$  est occupée et  $y_k = 0$  autrement, est connue pour  $k \in s$ , de sorte que l'estimation sans biais  $\bar{Y}_{\text{FUL}}$  définie par (3.2) peut être calculée et utilisée comme référence. Nous avons également calculé  $\bar{Y}_{\text{EXP}}$

$H_1$ ,  $H_2$  et  $H_3$  définis par (5.12).

Nous avons effectué la sélection de la façon suivante : à l'étape 0, le vecteur auxiliaire est le vecteur élémentaire  $x_k = 1$ , et l'estimateur est  $\bar{Y}_{\text{EXP}}$ . À l'étape 1, la valeur de l'indicateur est calculée pour chacune des 12 variables auxiliaires possibles ; la variable produisant la valeur la plus grande de l'indicateur est sélectionnée. À l'étape 2, la valeur de l'indicateur est calculée pour chacun des 11 vecteurs de dimension deux contenant la variable sélectionnée à l'étape 1 et l'une des variables restantes. La variable qui donne la plus grande valeur de l'indicateur est sélectionnée à l'étape 2, et ainsi de suite, aux étapes suivantes. Chaque nouvelle variable est jointe à celles déjà incluses dans le vecteur selon un schéma « côte à côte » (ou « + »). Par conséquent, il est fait abstraction des interactions. L'ordre de sélection est différent pour chaque indicateur.

Les valeurs de  $H_2$  et  $H_3$  qui déterminent la prochaine variable qui sera incluse dans le vecteur sont, par nécessité mathématique, plus grandes à chaque étape. Il n'en est toutefois pas ainsi pour  $H_1$ . À une certaine étape  $j$ , nous avons utilisé la règle consistant à inclure la variable  $x$  possédant la plus grande des valeurs calculées de  $H_1$ . Cette valeur peut être plus faible que la valeur de  $H_1$  qui a déterminé la variable sélectionnée à l'étape précédente,  $j - 1$ . La série de valeurs de  $H_1$  pour l'inclusion dans le vecteur augmente jusqu'à une certaine étape, puis commence à diminuer, comme l'illustre le tableau 9.1.

L'estimation sans biais est  $\bar{Y}_{\text{FUL}} = 4\,265$  ; l'estimation élémentaire est  $\bar{Y}_{\text{EXP}} = 4\,719$  (toutes les deux en milliers). Cela suggère un grand biais positif dans  $\bar{Y}_{\text{EXP}}$ , dont l'écart relatif (en %) par rapport à l'estimation sous réponse complète  $\bar{Y}_{\text{FUL}}$  est  $\text{ERC} = (\bar{Y}_{\text{EXP}} - \bar{Y}_{\text{FUL}}) / \bar{Y}_{\text{FUL}} \times 10^2 = 10,7$ . L'ajout de variables  $x$  catégoriques, une à une, dans le vecteur  $x$  modifiera successivement cet écart, quoique après avoir introduit quelques variables, la variation ne se produit pas toujours dans le sens d'une valeur plus faible. À chaque étape, nous avons calculé l'indicateur,  $\bar{Y}_{\text{CAL}}$  et  $\text{ERC} = (\bar{Y}_{\text{CAL}} - \bar{Y}_{\text{FUL}}) / \bar{Y}_{\text{FUL}} \times 10^2$ .

Le tableau 9.1 donne la sélection pas à pas avec l'indicateur  $H_1$  (le nombre de catégories est indiqué entre parenthèses pour chaque variable sélectionnée). La première variable sélectionnée est celle de la catégorie de revenu, qui produit une réduction importante de ERC, qui passe de 10,7 à 4,5. Les cinq sélections suivantes ont lieu pour des valeurs croissantes de  $H_1$ , et la valeur de ERC est réduite, mais d'une quantité successivement plus petite. L'étape 6, où la variable d'état matrimonial est sélectionnée, correspond à un point de virage, indiqué par la double ligne dans le tableau 9.1 : la valeur de  $H_1$  commence alors à diminuer, et



les valeurs inconnues de  $y$  qui caractérisent l'échantillon  $s$

ou la population  $U$ .

Considérons maintenant l'aspect (b) de l'ajustement,

c'est à dire l'ajustement hypothétique de la régression par les moindres carrés pondérés aux données  $(y_k, x_k)$  pour  $k \in s$ . Le vecteur des coefficients de régression sera  $\mathbf{B}^{x;s;d} = (\sum_s d_k x_k x_k')^{-1} \sum_s d_k x_k y_k$ , et les résidus  $e_{k;s;d} = y_k - \mathbf{x}_k' \mathbf{B}^{x;s;d}$  pour  $k \in s$  satisferont  $\sum_s d_k e_{k;s;d} = 0$ . Étant donné que  $\sum_r d_k m_k x_k / N = \bar{x}^{s;d}$  et  $\sum_r d_k m_k y_k / N = \bar{y}^{s;d}$  nous avons

$$\Delta_R = N^{-1}(\bar{Y}^{\text{CAL}} - \bar{Y}^{\text{FUL}}) = (1/N) \sum_r d_k m_k e_{k;s;d}. \quad (8.2)$$

Supposons que le modèle est « vrai pour l'échantillon  $s \rangle \rangle$ , avec un ajustement parfait, de sorte que  $e_{k;s;d} = 0$  pour tout  $k \in s$ . Alors, en vertu de (8.2), nous avons  $\Delta_R = 0$ , de sorte que l'estimateur corrigé de la non-réponse  $\bar{Y}^{\text{CAL}}$  concorde avec l'estimateur sans biais  $\bar{Y}^{\text{FUL}}$ . L'opinion que le biais est faible repose sur une hypothèse non vérifiable.

Penchons nous maintenant sur la question (ii) et

expliquons la relation essentiellement linéaire entre le biais de  $\bar{Y}^{\text{CAL}}$  et la valeur prévue de l'indicateur  $H_0 = \Delta_A / S_y = (\bar{Y}^{\text{EXP}} - \bar{Y}^{\text{CAL}}) / N S_y$ . Pour un résultat  $(s, r)$  donné, une variable  $y$  fixe et un vecteur  $\mathbf{x}$  fixe, nous avons

$$(\bar{Y}^{\text{CAL}} - Y) / N S_y = (\bar{Y}^{\text{EXP}} - Y) / N S_y - H_0.$$

Soit  $E^{pq}$  l'opérateur d'espérance par rapport à tous les résultats  $(s, r)$ , c'est à dire  $E^{pq}(\cdot) = E^p(E^q(\cdot|s))$ , où  $p(s)$  et  $q(r|s)$  sont, respectivement, le plan d'échantillonnage connu et la distribution des réponses inconnue. Nous écrivons  $\text{biais}(\bar{Y}^{\text{CAL}}) = E^{pq}(\bar{Y}^{\text{CAL}}) - Y$ ,  $\text{biais}(\bar{Y}^{\text{EXP}}) = E^{pq}(\bar{Y}^{\text{EXP}}) - Y$  et  $C = E^{pq}(N S_y)$ . En recourant au remplissage habituel en grand échantillon de la valeur prévue d'un ratio par le ratio des valeurs prévues, nous obtenons  $E^{pq}[(\bar{Y}^{\text{CAL}} - Y) / N S_y] \approx [E^{pq}(\bar{Y}^{\text{CAL}}) - Y] / E^{pq}(N S_y)$  et procédons de manière analogue pour  $\bar{Y}^{\text{EXP}}$ , de sorte que

$$\text{biais}(\bar{Y}^{\text{CAL}}) \approx \text{biais}(\bar{Y}^{\text{EXP}}) - C \times E(H_0). \quad (8.3)$$

Ici  $\text{biais}(\bar{Y}^{\text{EXP}})$  et  $C$  ne dépendent pas du choix du vecteur  $\mathbf{x}$ , tandis que  $\text{biais}(\bar{Y}^{\text{CAL}})$  et  $E(H_0)$  en dépendent. Donc, à mesure que le vecteur  $\mathbf{x}$  change,  $\text{biais}(\bar{Y}^{\text{CAL}})$  et  $E(H_0)$  sont reliés de manière essentiellement linéaire. Aucune forme particulière de  $p(s)$  et de  $q(r|s)$  ne doit être spécifiée pour que l'expression (8.3) soit vérifiée. Par conséquent, quand deux vecteurs auxiliaires,  $\mathbf{x}_{1k}$  et  $\mathbf{x}_{2k}$ , sont comparés, la différence de biais est, de façon étroitement approximative, proportionnelle à la variation de la valeur prévue de  $H_0$  :

$$\text{biais}(\bar{Y}^{\text{CAL}}(\mathbf{x}_{1k})) - \text{biais}(\bar{Y}^{\text{CAL}}(\mathbf{x}_{2k})) \approx -C(E_1 - E_2) \quad (8.4)$$

où  $E_i = E^{pq}(H_0(\mathbf{x}_{ik}))$  pour  $i = 1, 2$ . Les propriétés (8.3) et (8.4) sont validées par l'étude de Monte Carlo à la

section 10.

Notons que la formule (8.3) ne garantit pas que le biais de  $\bar{Y}^{\text{CAL}}$  basé sur un certain vecteur  $\mathbf{x}_k$  sera nul ou presque nul. Elle ne précise pas qu'une valeur comparativement grande de  $|\Delta_A|$  garantit un petit biais dans  $\bar{Y}^{\text{CAL}}$ . Ce que dit (8.3) est que  $\text{biais}(\bar{Y}^{\text{CAL}})$  est relié linéairement à l'espérance de l'indicateur  $H_0 = \Delta_A / S_y$ . Par conséquent, l'évaluation des vecteurs  $\mathbf{x}$  disponibles en fonction de l'indicateur  $H_0$  (ou de l'indicateur  $H_1 = |\Delta_A| / S_y$ ) est conforme à l'objectif de réduction du biais.

Passons maintenant au problème (iii) et commentons la méthode alternative de sélection des variables auxiliaires proposée par Schouten (2007). L'indicateur de ce dernier pour la sélection pas à pas des variables diffère des nôtres et ne produit habituellement pas exactement le même ensemble de variables. Dans une liste de, disons, 30 variables  $x$  catégoriques disponibles, les dix premières sélectionnées ne seront pas les mêmes que les dix retenues par nos indicateurs  $H_0$  à  $H_3$ . L'ordre dans lequel les variables sont sélectionnées ne sera pas nécessairement le même non plus. Nous avons comparé, dans certains de nos travaux empiriques, nos résultats au choix de variables réalisés selon la méthode de Schouten. Dans certains cas, nous avons constaté une congruence importante entre les deux séries de « dix premières variables » choisies selon les deux procédures.

Le meilleur moyen d'apprécier la différence entre les deux approches consiste à comparer leur contexte et leur démonstration mathématique. Nos indicateurs  $H_0$  et  $H_1$  reposent sur la notion de séparation (ou de distance), pour un résultat donné  $(s, r)$ , entre l'estimateur corrigé  $\bar{Y}^{\text{CAL}}$  et l'estimateur élémentaire,  $\bar{Y}^{\text{EXP}}$ , et sur l'idée que cette séparation augmentera habituellement quand le vecteur  $\mathbf{x}$  devient plus puissant. Le plan d'échantillonnage probabiliste est pris en considération et aucune hypothèse n'est formulée au sujet de la distribution des réponses.

Schouten s'appuie sur un argument en superpopulation et ne semble pas tenir compte des poids d'échantillonnage. Il trouve qu'une expression du biais prévu sous le modèle d'un estimateur de la moyenne de population est proportionnelle à la corrélation (au niveau de la population) entre la variable  $y$  et l'indicateur  $0 - 1$  de réponse. Il montre que cette corrélation (et, conséquemment, le biais) peut être bornée à l'intérieur d'un intervalle. En particulier, il considère l'estimateur par la régression généralisée et montre que son biais absolu maximum est égal à la largeur de l'intervalle de biais. Cette largeur dépend du vecteur réel inconnu de régression  $\beta$  pour la régression (au niveau de la population) de  $y$  en fonction de  $\mathbf{x}$ . Ce vecteur inconnu  $\beta$  est remplacé par une estimation basée sur les répondants, donc sujette à un certain biais, à cause de la non-réponse. Schouten met l'accent sur le fait qu'il n'est pas nécessaire d'émettre l'hypothèse que les données manquent au hasard



## 8. Commentaires : qualité de l'ajustement, propriétés du biais et procédures de sélection connexes

Trois problèmes sont examinées à la présente section : i) la relation entre le biais et la qualité de l'ajustement des régressions, ii) la relation linéaire entre la valeur prévue de  $\Delta_A = N^{-1}(X^{\text{EXP}} - Y^{\text{CAL}})$  et le biais de  $Y^{\text{CAL}}$  ou  $Y^{\text{FUL}}$ , et iii) la méthode alternative de sélection des variables auxiliaires proposée par Schouten (2007).

En ce qui concerne le problème (i), rappelons que l'écart total donné à la section 4 est  $\Delta_T = \Delta_A + \Delta_R$ , où  $\Delta_A$  est calculable, mais  $\Delta_T$  et  $\Delta_R$  ne le sont pas. S'il était calculable,  $N\Delta_R = Y^{\text{CAL}} - Y^{\text{FUL}}$  serait une estimation du biais de  $Y^{\text{CAL}}$  (et de celui de  $Y^{\text{FUL}}$ ). Une faible valeur de  $\Delta_R$  est souhaitable. La question qui se pose est celle de savoir si cela est réalisé quand le modèle  $Y_k = \beta'x_k + \varepsilon_k$  (pour un vecteur  $x_k$  donné) est bien ajusté aux données. Nous devons distinguer deux aspects : a) l'ajustement calculable aux données ( $Y_k, x_k$ ) observé pour  $k \in r$ , et b) l'ajustement hypothétique aux données ( $Y_k, x_k$ ) pour  $k \in s$ , certaines étant observées et d'autres pas.

Un bon ajustement pour les répondants,  $k \in r$ , ne garantit pas un petit  $\Delta_R$  : l'ajustement par les moindres carrés pondérés en utilisant les données observées ( $Y_k, x_k$ ) pour  $k \in r$  donne les résidus  $e^{k|r;d} = Y_k - x_k' \beta_{x^{r;d}}$ , calculables pour  $k \in r$ , avec la propriété  $\sum_r d^k e^{k|r;d} = 0$  (ici, la notation détaillée  $\beta_{x^{r;d}}$  spécifiée dans (2.8) est préférable à la notation simplifiée  $\beta_x$ ). Pour  $k \in s - r$ ,  $e^{k|r;d}$  n'est pas calculable ; il possède une moyenne non nulle inconnue  $\bar{e}^{s-r;d} = \sum_{s-r} d^k e^{k|r;d} / \sum_{s-r} d^k$ . Nous avons

$$\Delta_R = (Y^{\text{CAL}} - Y^{\text{FUL}}) / N = -(1 - P) \bar{e}^{s-r;d} \neq 0. \quad (8.1)$$

Que l'ajustement soit bon (petits résidus  $e^{k|r;d}$ ,  $R_{y,x}^2$  proche de un) ou mauvais (grands résidus  $e^{k|r;d}$ ,  $R_{y,x}^2$  proche de zéro), il se peut que l'écart  $\Delta_R$  donné par (8.1) soit grand et que  $Y^{\text{CAL}}$  soit loin d'être sans biais. Même si l'ajustement est parfait pour les répondants ( $e^{k|r;d} = 0$  pour tout  $k \in r$ , et  $R_{y,x}^2 = 1$ ), rien ne garantit que le biais sera faible.

Une inadéquation comparable affecte l'imputation basée sur les données fournies par les répondants. Si des imputations par la régression  $\hat{Y}_k = x_k' \beta_{x^{r;d}}$  sont utilisées pour remplacer les valeurs  $Y_k$  manquantes pour  $k \in s - r$ , l'estimateur imputé est donné par

$$Y_{\text{imp}}^{\text{CAL}} = \sum_r d^k Y_k + \sum_{s-r} d^k \hat{Y}_k.$$

Alors,  $Y_{\text{imp}}^{\text{CAL}} = Y^{\text{CAL}}$ , de sorte que l'exposition au biais est la même pour  $Y_{\text{imp}}^{\text{CAL}}$  que pour  $Y^{\text{CAL}}$ , ce qui est facile à comprendre : quand la non réponse cause une sélection biaisée des valeurs de  $Y$ , les valeurs imputées calculées en se basant sur cette sélection représenteront incorrectement

Le coefficient de corrélation entre  $y$  et le prédicteur univarié  $m$  est

$$R_{y,m}^2 = (C' \Sigma^{-1} C) / S_y^2. \quad (7.8)$$

Par conséquent, la proportion de  $S_y^2$  expliquée par  $m$  est

$$R_{y,m}^2 = -(D' \Sigma^{-1} C) / [(D' \Sigma^{-1} D)^{-1/2} \times S_y^2]. \quad (7.9)$$

Les proportions  $R_{y,x}^2$  et  $R_{y,m}^2$  satisfont  $R_{y,x}^2 \leq R_{y,m}^2 \leq 1$ .

*Preuve.* La preuve de (7.8) s'appuie sur la régression par les moindres carrés pondérés de  $y$  sur  $x$  ajustée sur  $r$ . Les résidus sont  $Y_k - \hat{Y}(x)_k$ , où  $\hat{Y}(x)_k = x_k' \beta_x$  avec  $\beta_x$  donné par (2.8). La décomposition est

$$\sum_r d^k (Y_k - \bar{Y}^{r;d})^2 = \sum_r d^k (\hat{Y}(x)_k - \bar{Y}^{r;d})^2 + \sum_r d^k (Y_k - \hat{Y}(x)_k)^2.$$

Le terme mixte est nul. Un développement du terme représentant la « variation expliquée » donne  $\sum_r d^k (\hat{Y}(x)_k - \bar{Y}^{r;d})^2 = (\sum_r d^k) C' \Sigma^{-1} C$ . Par conséquent, la proportion expliquée de la variance est  $R_x^2 = [\sum_r d^k (\hat{Y}(x)_k - \bar{Y}^{r;d})^2] / [(\sum_r d^k)^2 / S_y^2] = C' \Sigma^{-1} C / S_y^2$ , comme l'affirme (7.8). Pour démontrer (7.9), nous notons que la covariance, (5.7) peut s'écrire, avec l'aide de (7.5), sous la forme

$$\text{Cov}(y, m) = -\Delta_A / P = -D' \Sigma^{-1} C / P.$$

Il découle alors de (7.2) que  $R_{y,m} = \text{Cov}(y, m) / (S_y S_m)$  à pour expression (7.9). Les résidus de la régression (avec constante) de  $y$  sur la variable explicative univariée  $m$  sont  $\hat{Y}(m)_k = \bar{Y}^{r;d} + B^m(m_k - \bar{m}^{r;d})$  avec  $B^m = \text{Cov}(y, m) / S_m^2 = -P(D' \Sigma^{-1} C) / (D' \Sigma^{-1} D)$ . La proportion expliquée de la variance est  $\sum_r d^k (\hat{Y}(m)_k - \bar{Y}^{r;d})^2 / (\sum_r d^k S_y^2)$ , qui par développement donne l'expression de  $R_{y,m}^2$  correspondant à (7.10). Enfin,  $R_{y,x}^2 \leq R_{y,m}^2$  découle de l'inégalité de Cauchy Schwarz pour une forme bilinéaire :

$$(D' \Sigma^{-1} C)^2 \leq (D' \Sigma^{-1} D)(C' \Sigma^{-1} C).$$

L'inégalité  $R_{y,m}^2 \leq R_{y,x}^2 \leq 1$  peut également être déduite du fait que, parmi toutes les prédictions  $\hat{Y}_k = x_k' \beta$  linéaires dans le vecteur  $x$ , celles qui maximisent la variance expliquée sont  $\hat{Y}(x)_k = x_k' \beta_x$ , si bien que les prédictions  $\hat{Y}(m)_k$ , qui sont linéaires dans  $x_k$  par la voie de  $m_k$ , ne peuvent pas produire une plus grande variance expliquée que ce maximum.

Or, de (7.9), (7.2) et (7.5), il découle que  $-R_{y,m}^{\text{cv}} = D' \Sigma^{-1} C / S_y^2 = \Delta_A / S_y$ , comme l'affirme la formule (5.8). De surcroît, (7.7), (7.8) et (7.9) impliquent que  $-R_{y,m}^{\text{cv}} = \Delta_A$ , de sorte que le ratio des coefficients de corrélation  $r$  dans (5.9) est égal à l'angle  $\Delta$  défini par (7.7).

Pour chaque scénario, nous distinguons deux procédures :

*Procédure englobant tous les vecteurs* : Une liste de vecteurs  $\mathbf{x}$  possibles est dressée en se basant sur un jugement approprié. Nous calculons l'indicateur choisi pour chaque vecteur  $\mathbf{x}$  possible et choisissons celui qui donne la valeur la plus élevée de l'indicateur. Le vecteur  $\mathbf{x}$  résultant ne sera pas nécessairement le même pour  $H_1$  (qui a pour cible une variable  $y$  particulière) que pour  $H_3$  (qui recherche un compromis pour l'ensemble des variables  $y$  de l'enquête).

*Procédure pas à pas* (ou *Stepwise*) : Il existe un réservoir de variables  $x$  disponibles. Nous construisons le vecteur  $\mathbf{x}$  par sélection pas à pas ascendante (ou sélection pas à pas descendante) parmi les variables  $x$  disponibles, une variable à la fois, en nous basant sur les variations successives (si elles sont considérées suffisamment importantes) de la valeur de l'indicateur choisi pour déterminer l'inclusion (ou l'exclusion) d'une variable  $x$  donnée à une étape particulière. En général, les indicateurs  $H_1$ ,  $H_2$  et  $H_3$  ne produisent pas la même sélection de variables. Considérons deux vecteurs  $\mathbf{x}$ ,  $\mathbf{x}_{1k}$  et  $\mathbf{x}_{2k}$ , tels que  $\mathbf{x}_{2k}$  est composé de  $\mathbf{x}_{1k}$  et d'un vecteur supplémentaire  $\mathbf{x}_{2k} = (\mathbf{x}_{1k}', \mathbf{x}_{2k}')'$ . Le passage de  $\mathbf{x}_{1k}$  à  $\mathbf{x}_{2k}$  accroît la valeur de  $H_2$  et de  $H_3$ . À chaque étape d'une méthode de sélection ascendante, nous sélectionnons la variable qui produit l'accroissement le plus important de  $H_2$  ou de  $H_3$ . Toutefois, la transition ne garantit pas un accroissement de la valeur de l'indicateur le plus approprié,  $H_1$ . Néanmoins,  $H_1$  peut être utilisé dans la sélection pas à pas de la manière décrite à la section 9.

## 7. Démonstrations

Pour une variable  $y$  et un résultat  $(s, r)$  donné, nous recherchons un vecteur  $\mathbf{x}$  qui rend grand, en valeur absolue, le numérateur calculable  $\Delta_A = (\mathbf{x}_{r;d} - \mathbf{x}_{s;d})' \mathbf{B}_x$  du ratio des biais (4.3). À la présente section, nous prouvons les décompositions en facteurs  $\Delta_A / S_y^2 = -R_{y,m} \times \text{cv}_m^2 = F \times R_{y,x} \times \text{cv}_m^2$  données par (5.8) et (5.9). Nous commençons par noter que  $\text{cv}_m^2$  est une forme quadratique dans le vecteur qui contraste la moyenne de  $\mathbf{x}$  dans l'ensemble de réponses  $r$  avec la moyenne de  $\mathbf{x}$  dans l'échantillon  $s$ . Soit

$$\mathbf{D} = \mathbf{x}_{r;d} - \mathbf{x}_{s;d}; \quad \mathbf{\Sigma} = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' / \sum_r d_k. \quad (7.1)$$

Alors, avec  $P$  donné par (2.1),

$$\text{cv}_m^2 = P^2 \times S_y^2 = \mathbf{D}' \mathbf{\Sigma}^{-1} \mathbf{D}. \quad (7.2)$$

Cette expression découle de (5.3) et d'une conséquence de (2.3), à savoir

$$\mathbf{x}_{r;d}' \mathbf{\Sigma}^{-1} \mathbf{x}_{r;d} = \mathbf{x}_{s;d}' \mathbf{\Sigma}^{-1} \mathbf{x}_{s;d} = 1, \quad (7.3)$$

Le vecteur des covariances avec la variable étudiée  $y$  est donné par

$$\mathbf{C} = \left( \sum_r d_k (\mathbf{x}_k - \mathbf{x}_{r;d})(\mathbf{x}_k - \mathbf{x}_{r;d})' \right) / \left( \sum_r d_k \right), \quad (7.4)$$

Nous pouvons alors écrire  $\Delta_A$  sous une forme bilinéaire :

$$\Delta_A = \mathbf{D}' \mathbf{B}_x = \mathbf{D}' \mathbf{\Sigma}^{-1} \mathbf{C} \quad (7.5)$$

en utilisant le fait que  $\mathbf{D}' \mathbf{\Sigma}^{-1} \mathbf{x}_{r;d} = (\mathbf{x}_{r;d} - \mathbf{x}_{s;d})' \mathbf{\Sigma}^{-1} \mathbf{x}_{r;d} = 0$  en vertu de (7.3).

Une perspective utile de  $\Delta_A$  nous est fournie par l'interprétation géométrique de  $\mathbf{C}$  et  $\mathbf{D}$  figurant dans (7.5) en tant que vecteurs dans l'espace dont la dimension est celle de  $\mathbf{x}_k$ . Nous avons

$$\Delta_A = \Lambda (\mathbf{D}' \mathbf{\Sigma}^{-1} \mathbf{D})^{1/2} (\mathbf{C}' \mathbf{\Sigma}^{-1} \mathbf{C})^{1/2} \quad (7.6)$$

où

$$\Lambda = \frac{\mathbf{D}' \mathbf{\Sigma}^{-1} \mathbf{C}}{(\mathbf{D}' \mathbf{\Sigma}^{-1} \mathbf{D})^{1/2} (\mathbf{C}' \mathbf{\Sigma}^{-1} \mathbf{C})^{1/2}} \quad (7.7)$$

Pour une variable  $y$  particulière et un vecteur  $\mathbf{x}$  particulier, les quantités scalaires  $(\mathbf{D}' \mathbf{\Sigma}^{-1} \mathbf{D})^{1/2}$  et  $(\mathbf{C}' \mathbf{\Sigma}^{-1} \mathbf{C})^{1/2}$  représentent les longueurs vectorielles respectives de  $\mathbf{D}$  et  $\mathbf{C}$  (à la suite d'une transformation orthogonale basée sur les vecteurs propres et les valeurs propres de  $\mathbf{\Sigma}^{-1}$ ). La quantité scalaire  $\Lambda$  représente le cosinus de l'angle formé par  $\mathbf{D}$  (qui est indépendant de  $y$ ) et  $\mathbf{C}$  (qui dépend de  $y$ ) ; d'où  $-1 \leq \Lambda \leq 1$ .

Quand le vecteur auxiliaire  $\mathbf{x}_k$  peut s'agrandir par ajout de variables  $x$  supplémentaires, les deux longueurs vectorielles  $(\mathbf{D}' \mathbf{\Sigma}^{-1} \mathbf{D})^{1/2}$  et  $(\mathbf{C}' \mathbf{\Sigma}^{-1} \mathbf{C})^{1/2}$  augmentent. La variation de l'angle  $\Lambda$  peut se faire dans l'une ou l'autre direction ; Si  $|\Lambda|$  demeure approximativement constant, (7.6) montre que  $|\Delta_A|$  augmentera. Une deuxième perspective utile concernant  $\Delta_A$  découle de la décomposition de la variabilité totale de la variable étudiée  $y$ ,  $\sum_r d_k (y_k - \bar{y}_{r;d})^2 = (\sum_r d_k) S_y^2$ . Nous devons examiner l'ajustement de deux régressions, celle de  $y$  sur les vecteurs auxiliaires  $\mathbf{x}$ , et celle de  $y$  sur la variable dérivée  $m$  définie par (2.6). À chaque ajustement correspond une décomposition de  $S_y^2$  en une variation expliquée de  $y$  et une variation résiduelle de  $y$ . Les deux parties expliquées possèdent des liens importants avec le ratio des biais (4.3). Le résultat 7.1 résume les deux décompositions.

**Résultat 7.1.** Pour un résultat d'enquête donné  $(s, r)$ , soit  $\mathbf{D}$ ,  $\mathbf{\Sigma}$  et  $\mathbf{C}$  donnés par (7.1) et (7.4). Alors, la proportion de la variance  $S_y^2$  de  $y$  expliquée par la régression de  $y$  sur  $\mathbf{x}$  est



échantillons caractéristiques des enquêtes gouvernementales. Habituellement, la taille de l'échantillon est beaucoup plus grande que la dimension du vecteur  $\mathbf{x}$ . La variance des estimations est généralement faible comparativement au carré du biais. Néanmoins, pour les variables auxiliaires catégoriques, aucune taille de groupe « trop petite » ne devrait être permise. Il est recommandé que toutes les tailles de groupe valent au moins 30, voire au moins 50, afin d'éviter l'instabilité. Le croisement des variables catégoriques (pour permettre les interactions) comporte un certain risque qu'il existe des petits groupes. Il est préférable de caler sur les dénombrements marginaux, plutôt que sur les fréquences pour les cellules croisées de faible fréquence. Dans un certain nombre de pays, les nombreux registres administratifs existants constituent une riche source d'information auxiliaire, particulièrement pour les enquêtes auprès des particuliers et des ménages. Ces registres contiennent de nombreuses variables  $x$  possibles parmi lesquelles choisir. Un grand nombre de vecteurs  $\mathbf{x}$  différents peuvent donc être composés. Les indicateurs donnés par (5.12) offrent des outils de calcul pour obtenir un classement préférentiel des vecteurs  $\mathbf{x}$  possibles, l'objectif étant de réduire autant que possible le biais qui persiste dans l'estimateur par calage.

*Scénario 1 :* Concentrons nous sur une variable  $y$  particulière. Le biais persistant dans l'estimateur par calage dépend de la variable  $y$ ; certaines variables sont plus sujettes au biais que d'autres. Nous identifions une variable  $y$  particulière considérée comme étant très importante dans l'enquête, et nous cherchons à déterminer un vecteur  $\mathbf{x}$  qui réduit autant que possible le biais pour cette variable (s'il faut prendre en considération plus d'une variable  $y$ , un compromis doit être trouvé, ce que suggère le scénario 2 qui suit). Dans le présent exemple, nous utilisons l'indicateur  $H_1 = |\Delta_A|/S_y = |R_{y,m}| \times \text{cv}_m$  qui dépend de la variable  $y$  et choisissons le vecteur  $\mathbf{x}$  de façon à rendre  $H_1$  grand. Une alternative *ad hoc* consiste à utiliser l'indicateur  $H_2 = R_{y,x} \times \text{cv}_m$  et à s'efforcer de le rendre aussi grand que possible.

*Scénario 2 :* L'objectif est de déterminer un vecteur  $\mathbf{x}$  d'usage général, efficace pour toutes les variables  $y$  de l'enquête, ou pour la plupart d'entre elles. Cela suggère d'opter pour  $H_3 = \text{cv}_m$  comme indicateur de compromis et de choisir le vecteur  $\mathbf{x}$  qui maximise  $H_3$ . Dans ce même sens, Särndal et Lundström (2005, 2008) ont utilisé l'indicateur  $S_2^m = H_3^2/P_2^2$ . Ils ont montré que la variable dérivée  $m_k$  dans (2.6) peut être considérée comme un prédicteur de l'inverse de la probabilité de réponse inconnue et que choisir le vecteur  $\mathbf{x}$  de manière à rendre  $S_2^m$  grand signale une réduction du biais dans l'estimateur par calage, indépendamment de la variable  $y$ .

l'ordre des milliers. Il reste à savoir si cette correction a éliminé ou non la plupart du biais dû à la non réponse. Il découle de (5.8) que  $0 \leq |\Delta_A|/S_y \leq \text{cv}_m$ , quelle que soit la variable  $y$ . L'inégalité  $|\Delta_A|/S_y \leq R_{y,x} \times \text{cv}_m$  est meilleure, mais elle dépend de la variable  $y$ . En outre, si le ratio de corrélation  $F$  demeure approximativement constant quand le vecteur  $\mathbf{x}$  change, de sorte que  $F \approx F_0$ , alors  $|\Delta_A|/S_y \approx |F_0| \times R_{y,x} \times \text{cv}_m$ . Bien qu'il soit calculable pour tout vecteur  $\mathbf{x}$  et tout résultat  $(s, r)$ ,  $\Delta_A$  ne révèle pas la valeur du ratio des biais. Cependant,  $\Delta_A$  suggère des outils de calcul, appelés indicateurs, pour comparer divers vecteurs  $\mathbf{x}$ . Partant de (5.8), posons que

$$H_0 = \Delta_A/S_y = -R_{y,m} \times \text{cv}_m. \quad (5.11)$$

Comme le prouve la théorie exposée à la section 8 et les travaux empiriques présentés à la section 10, sur une longue série de résultats  $(s, r)$ , la moyenne de  $H_0$  concorde avec l'écart moyen  $\bar{Y}^{\text{CAL}} - Y$  (qui mesure le biais de  $\bar{Y}^{\text{CAL}}$ ) de manière presque parfaitement linéaire quand le vecteur  $\mathbf{x}$  change. Il en est ainsi indépendamment de la distribution des réponses qui génère  $r$  à partir de  $s$ . Puisque  $H_0$  peut prendre l'un ou l'autre signe, il est pratique de travailler avec sa valeur absolue désignée par  $H_1$ ; en plus, nous considérons deux autres indicateurs,  $H_2$  et  $H_3$ , inspirés de (5.9) et (5.10) :

$$H_1 = |\Delta_A|/S_y = |R_{y,m}| \times \text{cv}_m; \quad H_2 = R_{y,x} \times \text{cv}_m; \quad H_3 = \text{cv}_m. \quad (5.12)$$

Nos principales alternatives sont  $H_1$  et  $H_3$ . De celles-ci,  $H_1$  est motivé par son lien direct avec  $\Delta_A$ , que nous voulons rendre plus grand, pour une variable  $y$  donnée. Une bonne raison de considérer  $H_3$  est son indépendance à l'égard de toutes les variables  $y$  de l'enquête. L'indicateur  $H_2$  est une alternative *ad hoc*; bien qu'il contienne un concept connu, le coefficient de corrélation multiple  $R_{y,x}$ , il est moins approprié que  $H_1$ , parce que le ratio des coefficients de corrélation  $F = -R_{y,m}/R_{y,x}$  peut varier considérablement d'un vecteur  $\mathbf{x}$  à l'autre. Aussi bien  $H_2$  que  $H_3$  augmentent quand d'autres variables  $x$  sont ajoutées au vecteur  $\mathbf{x}$ , ce qui n'est généralement pas vérifié pour  $H_1$ . L'utilisation de ces indicateurs est illustrée aux sections 9 et 10 décrivant les travaux empiriques.

## 6. Classement préférentiel des vecteurs auxiliaires

Les méthodes décrites dans le présent article sont destinées à être utilisées principalement avec les grands



## 5. Expression de l'écart expliqué

L'unité répondante  $k$  reçoit le poids  $d^k m_k$  dans l'estimateur  $\hat{Y}_{CAL} = \sum^r d^k m_k y_k$ . Le facteur de correction de la non réponse  $m_k = (\sum^s d^k x_k) (\sum^r d^k x_k x_k')^{-1} x_k$  accroît le poids de sondage  $d^k$ . Nous pouvons considérer  $m_k$  comme la valeur d'une variable dérivée, définie pour un résultat ( $r$ ,  $s$ ) et un choix de  $x_k$ , particuliers, indépendante de toutes les variables  $y$  d'intérêt, et calculable pour  $k \in s$  (mais utilisée dans  $\hat{Y}_{CAL}$  uniquement pour  $k \in r$ ). En utilisant (2.3), nous avons

$$\sum^r d^k m_k x_k = \sum^s d^k x_k x_k'; \quad \sum^r d^k m_k = \sum^s d^k; \quad (5.1)$$

Deux moyennes pondérées sont nécessaires :

$$\bar{m}_{r,d} = \frac{\sum^r d^k m_k}{\sum^r d^k} = \frac{\sum^s d^k}{\sum^s d^k} = \frac{1}{P}; \quad \bar{m}_{s,d} = \frac{\sum^s d^k}{\sum^s d^k} \quad (5.2)$$

où  $P$  est le taux de réponse (2.1). Donc, le facteur de correction moyen dans  $\hat{Y}_{CAL} = \sum^r d^k m_k y_k$  est  $1/P$ , quel que soit le choix du vecteur  $x$ . La capacité que possède un vecteur  $x$  choisi de réduire efficacement ou non le biais dépendra des moments d'ordre plus élevé de  $m_k$ . La

variance pondérée de  $m_k$  est donnée par

$$S_{m_{r,d}}^2 = S_{m_{s,d}}^2 = \sum^r d^k (m_k - \bar{m}_{r,d})^2 / \sum^r d^k \quad (5.3)$$

Nous utiliserons la notation simplifiée  $S_{m_{r,d}}^2$ . Un développement de (5.3), et l'utilisation de (5.1) et de (5.2) donnent

$$S_{m_{r,d}}^2 = \bar{m}_{r,d} (\bar{m}_{s,d} - \bar{m}_{r,d}). \quad (5.4)$$

Le coefficient de variation de  $m_k$  est

$$cv_m = \frac{S_{m_{r,d}}}{\bar{m}_{r,d}} = \sqrt{\frac{\bar{m}_{s,d}}{\bar{m}_{r,d}} - 1}. \quad (5.5)$$

La variance pondérée de la variable étudiée  $y$  est donnée

par

$$S_y^2 = S_{y|U}^2 = \sum^r d^k (y_k - \bar{y}_{r,d})^2 / \sum^r d^k \quad (5.6)$$

(quand les probabilités de réponse ne sont pas toutes égales,  $S_y^2 = S_{y|U}^2$  n'est pas sans biais pour la variance de population  $S_{y|U}^2$ , mais il ne s'agit pas d'un problème pour les dérivations qui suivent). Nous avons besoin de la covariance

$$\text{Cov}(y, m) = \text{Cov}(y, m)_{r,d} =$$

$$\frac{1}{I} \sum^r d^k (m_k - \bar{m}_{r,d})(y_k - \bar{y}_{r,d}) \quad (5.7)$$

et du coefficient de corrélation  $R_{y,m} = \text{Cov}(y, m) / (S_y S_m)$ , satisfaisant  $-1 \leq R_{y,m} \leq 1$ . L'écart  $\Delta_A = (\bar{x}_{r,d} - \bar{x}_{s,d})' B_x$  est une composante essentielle du ratio des biais (4.3). Nous recherchons un vecteur  $x$  qui rend  $\Delta_A$  grand. Les facteurs qui déterminent  $\Delta_A$  figurent dans les expressions (5.8) à (5.10). Les outils de calcul (indicateurs) destinés à faciliter la recherche des variables  $x$  efficaces sont donnés en (5.11) et (5.12). Leur dérivation par l'algèbre linéaire est reportée à la section 7, que peuvent omettre les lecteurs qui s'intéressent surtout à l'utilisation pratique de ces outils pour trouver les variables  $x$ , comme l'illustre empiriquement les sections 9 et 10. Nous pouvons décomposer  $\Delta_A / S_y$  en facteurs comme il suit

$$\Delta_A / S_y = -R_{y,m} \times cv_m. \quad (5.8)$$

Deux facteurs multiplicatifs simples déterminent  $\Delta_A / S_y$  : le coefficient de variation  $cv_m$ , qui est indépendant de  $y_k$  et calculé sur le vecteur  $x_k$  connu uniquement, et le coefficient de corrélation (positif ou négatif)  $R_{y,m}$ . Une autre décomposition en facteurs basée sur des concepts simples est

$$\Delta_A / S_y = F \times R_{y,x} \times cv_m \quad (5.9)$$

où  $R_{y,x} = \sqrt{R_{y,x}^2}$  est le coefficient de corrélation multiple entre  $y$  et  $x$ ,  $R_{y,x}^2$  est la proportion de la variance  $S_y^2$  de  $y$  expliquée par le prédicteur  $x$ , et  $F = -R_{y,m} / R_{y,x}$  (la formule (7.8) donne l'expression précise pour  $R_{y,x}^2$ ). Comme nous le montrons également à la section 7,  $|R_{y,m}| \leq R_{y,x}$  pour tout vecteur  $x$  et toute variable  $y$ ; par conséquent,  $-1 \leq F \leq 1$ .

Dans (5.8) et (5.9),  $cv_m$  et  $R_{y,x}$  sont des termes non négatifs, tandis que  $R_{y,m}$  et  $F$  peuvent prendre n'importe quel signe (ou éventuellement être nuls). Donc,

$$|\Delta_A| / S_y = |R_{y,m}| \times cv_m = |F| \times R_{y,x} \times cv_m. \quad (5.10)$$

Les termes  $S_y$ ,  $cv_m$ ,  $R_{y,x}$ ,  $R_{y,m}$  et  $F$  se calculent tous facilement d'après les données d'enquête. Les termes  $cv_m$  et  $R_{y,x}$  augmentent tous deux (ou demeurent éventuellement invariables) quand des variables  $x$  supplémentaires sont ajoutées au vecteur  $x$ ;  $R_{y,m}$  ne possède pas cette propriété.

Illustrons cela à l'aide de chiffres assez habituels. Si  $F = 0,5$ ;  $R_{y,x} = 0,6$  et  $cv_m = 0,4$ , alors  $\Delta_A / S_y = 0,12$ , ce qui implique que  $\hat{Y}_{CAL} / N = \hat{Y}_{EXP} / N - 0,12 \times S_y$ . Autrement dit, la moyenne de  $y$  estimée  $\hat{Y}_{CAL} / N$  a été rajustée à la baisse d'un facteur égal à 0,12 écart-type par rapport à l'estimation élémentaire  $\hat{Y}_{EXP} / N$ . La correction peut être importante comparativement à l'écart-type de la moyenne estimée de  $y$ , surtout quand la taille de l'échantillon est de

En résumé, pour un résultat donné  $(s, r)$  et une variable  $y$  donnée, les trois écarts possèdent les caractéristiques suivantes : i)  $\Delta_T = \bar{y}_{r;d} - \bar{y}_{s;d}$  est une valeur constante inconnue, qui dépend des valeurs de  $y$  non observées ainsi qu'observées, ii)  $\Delta_A$  est calculable, sa valeur dépend de  $y_k$  pour  $k \in r$  et des valeurs de  $x_k$  pour  $k \in s$  pour le vecteur  $x$  choisi, iii)  $\Delta_R$  ne peut pas être calculé, sa valeur dépend des valeurs non observées  $y_k$ , et de  $x_k$  pour  $k \in s$ .

Afin de suivre l'évolution des estimations quand le vecteur  $x$  s'améliore, considérons un résultat donné  $(s, r)$ . L'écart  $\Delta_T$  peut avoir n'importe quel signe. Supposons que  $\Delta_T > 0$ , qui indique un biais positif dans  $\tilde{Y}^{\text{EXP}}$ , comme dans le cas où de grandes unités manifestent une plus grande propension à répondre que les petites. À mesure que le vecteur  $x$  dans  $\tilde{Y}^{\text{CAL}}$  devient plus puissant grâce à l'inclusion d'un nombre croissant de variables  $x$ ,  $\Delta_A$  a tendance à croître et à s'écarter de zéro et, idéalement, s'approchera de  $\Delta_T$ , indiquant une proximité souhaitée de  $\tilde{Y}^{\text{CAL}}$  et de l'estimation  $\tilde{Y}^{\text{FUL}}$  sans biais. Aussi longtemps que le vecteur  $x$  demeure relativement faible, il est vraisemblable que l'inégalité  $\Delta_A < \Delta_T$  soit vérifiée. À mesure que la puissance du vecteur  $x$  s'accroît,  $\Delta_A$  se rapproche de l'écart fixe  $\Delta_T$ , signe que le biais devient presque nul. Il pourrait même « aller au-delà », de sorte qu'un « surajustement »,  $\Delta_A > \Delta_T$ , se produira. Cette situation n'est pas nuisible, quoique  $\Delta_R = \Delta_T - \Delta_A$  est alors négatif, il est ordinairement petit (l'analyste ne peut travailler qu'avec  $\Delta_A$  ; il sait pas quand  $\Delta_A$  et  $\Delta_T$  sont proches, ni si le surajustement  $\Delta_A > \Delta_T$  a eu lieu). La simulation décrite à la section 10 illustre ces points. Si  $\Delta_T < 0$ , ces tendances sont inversées.

La forme de (4.3) peut évoquer un argument susceptible d'être incorrect. Supposons que l'on ait proposé un vecteur  $x_k$ , contenant des variables considérées comme efficaces et que l'on ait émis l'hypothèse que  $y_k = \beta'x_k + \varepsilon_k$ , où  $\varepsilon_k$  est un petit résidu. Alors,  $\bar{y}_{r;d} \approx (\bar{x}_{r;d})'\beta_x \approx (\bar{x}_{r;d} - \bar{x}_{s;d})'\beta_x$  et conséquemment, le ratio des biais  $\approx 0$ , ce qui communique le message, souvent faux, que le vecteur postulé  $x_k$  est efficace. L'une des faiblesses de l'argument découle du fait bien connu que la non réponse (à moins qu'elle soit entièrement aléatoire) induira un biais dans  $\tilde{Y}^{\text{CAL}}$  pour un vecteur de régression qui décrit la relation de  $y$  en fonction de  $x$  dans la population. D'autres commentaires à ce sujet sont donnés à la section 8.

Enfin, il faut tenir compte du fait qu'en pratique, une enquête porte habituellement sur de nombreuses variables  $y$ . À chaque variable  $y$  correspond un estimateur par calage et un ratio des biais donné par (4.3). Le vecteur  $x$  idéal est celui qui serait capable de contrôler le biais dans tous ces estimateurs, ce qui est habituellement impossible sans compromis, comme nous en discutons plus loin.

$\tilde{Y}^{\text{CAL}} - \tilde{Y}^{\text{FUL}}$ , parmi lesquels seul celui du milieu est calculable. L'« écart total » inconnu,  $\tilde{Y}^{\text{EXP}} - \tilde{Y}^{\text{FUL}}$ , peut être décomposé en un « écart expliqué » (par le vecteur  $x$  choisi) plus un « écart restant » :

$$(4.1) \quad \tilde{Y}^{\text{EXP}} - \tilde{Y}^{\text{FUL}} = (\tilde{Y}^{\text{EXP}} - \tilde{Y}^{\text{CAL}}) + (\tilde{Y}^{\text{CAL}} - \tilde{Y}^{\text{FUL}}).$$

S'ils étaient calculables, l'écart  $\tilde{Y}^{\text{CAL}} - \tilde{Y}^{\text{FUL}}$  serait particulièrement intéressant en tant qu'estimation du biais persistant dans  $\tilde{Y}^{\text{CAL}}$  (et dans  $\tilde{Y}^{\text{CAL}}$ ), tandis que l'écart  $\tilde{Y}^{\text{EXP}} - \tilde{Y}^{\text{FUL}}$  estimerait le biais habituellement beaucoup plus grand de l'estimation repère,  $\tilde{Y}^{\text{EXP}}$ . Le ratio des biais pour un résultat donné  $(s, r)$  détermine le biais estimé de  $\tilde{Y}^{\text{CAL}}$  par rapport à celui de  $\tilde{Y}^{\text{EXP}}$  :

$$(4.2) \quad \text{ratio des biais} = \frac{\tilde{Y}^{\text{CAL}} - \tilde{Y}^{\text{FUL}}}{\tilde{Y}^{\text{EXP}} - \tilde{Y}^{\text{FUL}}}.$$

Nous normalisons les trois écarts en utilisant comme facteur la taille estimée de la population  $N = \sum_s d_k$  et utilisons la notation  $\Delta_T = \Delta_A + \Delta_R$ , où  $T$  signifie « total »,  $A$  signifie « expliqué (*accounted for*) » et  $R$  signifie « restant ». En notant que  $\sum_r d_k (y_k - x_k' \beta_x) = 0$ , nous avons

$$\Delta_T = N^{-1}(\tilde{Y}^{\text{EXP}} - \tilde{Y}^{\text{FUL}}) = \bar{y}_{r;d} - \bar{y}_{s;d};$$

$$\Delta_R = N^{-1}(\tilde{Y}^{\text{CAL}} - \tilde{Y}^{\text{FUL}}) = \bar{x}_{r;d}'\beta_x - \bar{y}_{s;d}$$

$$\Delta_A = N^{-1}(\tilde{Y}^{\text{EXP}} - \tilde{Y}^{\text{CAL}}) = (\bar{x}_{r;d} - \bar{x}_{s;d})'\beta_x$$

où  $\bar{x}_{s;d} = \sum_s d_k x_k / \sum_s d_k$ ,  $\bar{x}_{r;d} = \sum_r d_k x_k / \sum_r d_k$ , et  $\bar{y}_{r;d}$  et  $\bar{y}_{s;d}$  sont les moyennes définies de manière analogue pour la variable  $y$ . Alors, (4.2) prend la forme :

$$(4.3) \quad \text{ratio des biais} = \frac{\Delta_T}{\Delta_R} = 1 - \frac{\Delta_T}{\Delta_A} = 1 - \frac{(\bar{x}_{r;d} - \bar{x}_{s;d})'\beta_x}{\bar{y}_{r;d} - \bar{y}_{s;d}}.$$

Pour le vecteur élémentaire  $x_k = 1$ , le ratio des biais = 1. Idéalement, nous voulons que le vecteur auxiliaire  $x_k$  pour  $\tilde{Y}^{\text{CAL}}$  donne un ratio des biais  $\approx 0$ . Pour un résultat donné  $(s, r)$  et une variable  $y$  donnée, nous progressons dans cette direction en trouvant un vecteur  $x$  qui rend le numérateur calculable  $\Delta_A = (\bar{x}_{r;d} - \bar{x}_{s;d})'\beta_x$  grand (en valeur absolue), ce qui est réalisable. Cependant, quel que soit le vecteur  $x$  que nous choisissons finalement, le biais restant dans  $\tilde{Y}^{\text{CAL}}$  est inconnu. Même en utilisant le vecteur  $x$  le meilleur possible, un biais important peut persister. Nous avons donc tenté de trouver la meilleure solution possible, dans des conditions peut-être défavorables.



$$\hat{Y}^{\text{CAL}} = \sum_{\mathbf{x}_k} w_k \mathbf{y}_k \quad (2.4)$$

avec  $w_k = d_k^{-1} \{1 + (\mathbf{X} - \sum_{\mathbf{x}_k} d_k \mathbf{x}_k \mathbf{x}_k') (\sum_{\mathbf{x}_k} d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k\}$ . Les poids  $w_k$  sont calés sur deux types d'information :

$\sum_{\mathbf{x}_k} w_k \mathbf{x}_k = \mathbf{X}$ , qui implique que  $\sum_{\mathbf{x}_k} w_k \mathbf{x}_k' = \sum_{\mathbf{x}_k} \mathbf{x}_k' w_k$  et  $\sum_{\mathbf{x}_k} w_k \mathbf{x}_k' \mathbf{x}_k = \sum_{\mathbf{x}_k} d_k \mathbf{x}_k' \mathbf{x}_k$ . Nous supposons tout au long de l'exposé que la matrice symétrique  $\sum_{\mathbf{x}_k} d_k \mathbf{x}_k \mathbf{x}_k'$  est non singulière (pour des raisons de calcul, il est prudent d'imposer une contrainte plus forte : la matrice ne doit pas être mal conditionnée, ou presque singulière). Étant donné (2.3), nous avons  $\hat{Y}^{\text{CAL}} = \sum_{\mathbf{x}_k} w_k \mathbf{y}_k$  avec les poids  $w_k = d_k \mathbf{v}_k$ , où  $\mathbf{v}_k = \mathbf{X}' (\sum_{\mathbf{x}_k} d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$ . Les poids satisfont  $\sum_{\mathbf{x}_k} d_k \mathbf{v}_k \mathbf{x}_k = \mathbf{X}$ , où  $\mathbf{X}$  contient l'une des composantes de (2.2), ou les deux.

Un estimateur par calage étroitement apparenté est celui basé sur le même vecteur à deux composantes  $\mathbf{x}_k$ , mais avec le calage uniquement au niveau de l'échantillonnage :

$$\hat{Y}^{\text{CAL}} = \sum_{\mathbf{x}_k} d_k m_k \mathbf{y}_k \quad (2.5)$$

où

$$m_k = \left( \sum_{\mathbf{x}_k} d_k \mathbf{x}_k \right)' \left( \sum_{\mathbf{x}_k} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_k \quad (2.6)$$

L'équation de calage se lit alors  $\sum_{\mathbf{x}_k} d_k m_k \mathbf{x}_k = \sum_{\mathbf{x}_k} d_k \mathbf{x}_k$ , où  $\mathbf{x}_k$  comprend les deux composantes données en (2.2). Le

vecteur auxiliaire  $\mathbf{x}_k$  sert à réaliser deux objectifs : obtenir une faible variance et un faible biais de non réponse. Du point de vue de la variance uniquement,  $\hat{Y}^{\text{CAL}}$  est habituellement préféré à  $\hat{Y}^{\text{FUL}}$ , parce que le premier bénéficie de l'apport d'un total de population connu  $\sum_{\mathbf{x}_k} \mathbf{x}_k'$ . Toutefois, comme la présente étude porte sur le biais, cela revient pour ainsi dire au même d'utiliser  $\hat{Y}^{\text{CAL}}$  ou  $\hat{Y}^{\text{FUL}}$ , et nous nous concentrons sur le second. Sous des conditions libérales, la différence entre le biais de  $N^{-1} \hat{Y}^{\text{CAL}}$  et celui de  $N^{-1} \hat{Y}^{\text{FUL}}$  est d'ordre  $n^{-1}$ , et a donc peu de conséquences pratiques, même pour des tailles d'échantillon  $n$  modestes, comme il est discuté, par exemple, dans Särndal et Lundström (2005).

L'estimateur (2.5) peut aussi s'exprimer sous la forme

$$\hat{Y}^{\text{CAL}} = \left( \sum_{\mathbf{x}_k} d_k \mathbf{x}_k \right)' \mathbf{B}^* \quad (2.7)$$

où

$$\mathbf{B}^* = \mathbf{B}_{\mathbf{x}^{\text{CAL}}} = \left( \sum_{\mathbf{x}_k} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{\mathbf{x}_k} d_k \mathbf{x}_k \mathbf{y}_k \quad (2.8)$$

est le vecteur des coefficients de régression résultant de l'ajustement par les moindres carrés (pondérés par  $d_k$ ) sur les données  $(\mathbf{y}_k, \mathbf{x}_k)$  pour  $k \in r$ .

Remarque concernant la notation : Lorsque cela est nécessaire pour insister sur un fait, un symbole possède deux indices séparés par un point virgule. Le premier indique l'ensemble d'unités sur lequel la quantité est

### 3. Points de référence

Le choix de vecteur le plus élémentaire est la valeur unitaire constante,  $\mathbf{x}_k = 1$  pour tout  $k$ . Bien que ce vecteur soit inefficace en ce qui concerne la réduction du biais de non réponse, il sert de valeur repère. Alors,  $m_k = 1/P$  pour tout  $k$ , où  $P$  est le taux de réponse à l'enquête (2.1) et  $\hat{Y}^{\text{CAL}}$  est l'estimateur à facteur d'extension (*expansion estimator*) :

$$\hat{Y}^{\text{EXP}} = (1/P) \sum_{\mathbf{x}_k} d_k \mathbf{y}_k = N \bar{y}_{r;d} \quad (3.1)$$

où  $N = \sum_{\mathbf{x}_k} d_k$  est sans biais sous le plan pour la taille de population  $N$ . Le biais de  $\hat{Y}^{\text{EXP}}$  peut être important.

À l'autre extrémité du spectre de biais se trouvent les estimateurs sans biais, ou presque sans biais, qui peuvent être obtenus sous réponse complète, quand  $r = s$ . Il s'agit d'estimateurs hypothétiques, non calculables en présence de non réponse. Parmi eux figure l'estimateur GREG avec les poids calés sur le total de population connu  $\sum_{\mathbf{x}_k} \mathbf{x}_k'$ .

$$\hat{Y}^{\text{FUL}} = \sum_{\mathbf{x}_k} d_k g_k \mathbf{y}_k$$

où  $g_k = 1 + (\sum_{\mathbf{x}_k} \mathbf{x}_k' - \sum_{\mathbf{x}_k} d_k \mathbf{x}_k') (\sum_{\mathbf{x}_k} d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k'$  et FUL désigne la réponse complète (*full response*). L'estimateur HT sans biais (obtenu quand  $g_k = 1$  pour tout  $k$ ) s'écrit

$$\hat{Y}^{\text{FUL}} = \sum_{\mathbf{x}_k} d_k \mathbf{y}_k = N \bar{y}_{r;d} \quad (3.2)$$

Il ne tient pas compte de l'information  $\sum_{\mathbf{x}_k} \mathbf{x}_k'$ , qui peut être importante pour la réduction de la variance. Cependant, pour l'étude du biais exposée ici, nous pouvons utiliser indifféremment  $\hat{Y}^{\text{FUL}}$  ou  $\hat{Y}^{\text{HT}}$ . La différence de biais entre les deux a peu d'importance, même pour des tailles d'échantillon modestes. Nous pouvons donc nous con-

centrer sur  $\hat{Y}^{\text{FUL}}$ .

### 4. Le ratio des biais

Pour un résultat donné ( $s, r$ ), considérons les estimations  $\hat{Y}^{\text{CAL}}$ ,  $\hat{Y}^{\text{EXP}}$  et  $\hat{Y}^{\text{FUL}}$ , données par (2.5), (3.1) et (3.2), comme trois points sur un axe horizontal. Les estimations  $\hat{Y}^{\text{EXP}}$  (produite par le vecteur élémentaire  $\mathbf{x}_k = 1$ ) et  $\hat{Y}^{\text{CAL}}$  (produite par un meilleur vecteur  $\mathbf{x}$ ) sont calculables, mais biaisées. À mesure que s'améliore le vecteur  $\mathbf{x}$ ,  $\hat{Y}^{\text{CAL}}$  s'écarte de  $\hat{Y}^{\text{EXP}}$  et peut se rapprocher de l'estimation  $\hat{Y}^{\text{FUL}}$  sans biais idéale, mais non réalisée. Nous considérons par conséquent trois écarts :  $\hat{Y}^{\text{EXP}} - \hat{Y}^{\text{FUL}}$ ,  $\hat{Y}^{\text{EXP}} - \hat{Y}^{\text{CAL}}$  et



*auxiliaire d'échantillon* est transmise par  $\mathbf{x}_0^k$ , une valeur vectorielle connue (observée) pour chaque unité  $k \in s$ ; le total  $\sum_U \mathbf{x}_0^k$  est inconnu, mais est estimé sans biais par  $\sum_s d_k^k \mathbf{x}_0^k$ . La valeur vectorielle auxiliaire combinant les deux types de vecteurs est désignée  $\mathbf{x}_k$ . Ce vecteur et l'information qui y est associée sont

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_*^k \\ \mathbf{x}_0^k \end{pmatrix}; \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_*^k \\ \sum_s d_k^k \mathbf{x}_0^k \end{pmatrix}. \quad (2.2)$$

Le vecteur  $(y_k, \mathbf{x}_k, \pi_k)$  est lié à la  $k^e$  unité. Ici,  $\pi_k$  est connue pour tout  $k \in U$ ,  $y_k$  pour tout  $k \in r$ , la composante  $\mathbf{x}_*^k$  de  $\mathbf{x}_k$  fournit l'information de population et la composante  $\mathbf{x}_0^k$  de  $\mathbf{x}_k$  fournit l'information d'échantillon. De nombreux vecteurs  $\mathbf{x}$  peuvent être formés à l'aide des variables extraites des registres administratifs, des données sur les processus d'enquête ou d'autres sources. Parmi tous les vecteurs à notre disposition, nous voulons identifier celui qui est le plus susceptible de réduire le biais de non réponse, si ce n'est pas à une valeur nulle, au moins à une valeur quasi nulle.

Nous considérons les vecteurs ayant la propriété qu'il existe un vecteur non nul constant  $\boldsymbol{\mu}$  tel que

$$\boldsymbol{\mu}' \mathbf{x}_k = 1 \text{ pour tout } k \in U \quad (2.3)$$

« Constant » signifie que  $\boldsymbol{\mu} \neq \mathbf{0}$  ne dépend pas de  $k$ , ni de  $s$  ou de  $r$ . La condition (2.3) simplifie les démonstrations mathématiques et ne contraint pas sévèrement  $\mathbf{x}_k$ . En fait, la plupart des vecteurs  $\mathbf{x}$  utiles en pratique sont couverts. À titre d'exemple, mentionnons : 1)  $\mathbf{x}_k = (1, x_k)'$ , où  $x_k$  est la valeur, pour l'unité  $k$ , d'une variable auxiliaire continue  $x$ ; 2) le vecteur représentant une variable  $x$  catégorique avec  $J$  classes mutuellement exclusives et exhaustives,  $\mathbf{x}_k = \gamma_k = (\gamma_{jk}, \dots, \gamma_{Jk})'$ , où  $\gamma_{jk} = 1$  si  $k$  appartient au groupe  $j$ , et  $\gamma_{jk} = 0$  autrement,  $j = 1, 2, \dots, J$ ; 3) le vecteur  $\mathbf{x}_k$  utilisé pour codifier deux variables catégoriques, la dimension de  $\mathbf{x}_k$  étant  $J_1 + J_2 - 1$ , où  $J_1$  et  $J_2$  sont les nombres respectifs de classes, et où le « moins un » a pour but d'éviter une singularité dans le calcul des poids calés sur les deux matrices des dénombrements marginaux; 4) l'extension de 3) à plus de deux variables catégoriques. Les vecteurs de type 3) et 4) sont particulièrement importants pour la production de statistiques par les organismes statistiques (le choix  $\mathbf{x}_k = x_k$ , non couvert par (2.3), mène à l'estimateur de la non réponse par le ratio qui a la réputation d'être habituellement un mauvais choix pour contrôler le biais de non réponse comparativement à  $\mathbf{x}_k = (1, x_k)'$ , si bien qu'exclure l'estimateur par le ratio ne constitue pas une grande perte).

L'estimateur par calage de  $Y = \sum_U y_k$ , calculé sur les données  $y_k$  pour  $k \in r$ , est

Särndal et Lundström (2005, 2008) proposent ce genre d'indicateurs, tandis que Schouten (2007) adopte une perspective différente pour justifier un indicateur. Un article à consulter à ce sujet est celui de Schouten, Cobben et Bethlehem (2009).

Le présent article comporte quatre volets. Aux sections 2 à 4, nous exposons le contexte général de l'estimation en présence de non réponse. Aux sections 5 et 6, nous présentons les indicateurs pour le classement préférentiel des vecteurs  $\mathbf{x}$  et nous discutons de leur calcul. Aux sections 7 et 8, nous présentons l'algèbre linéaire qui sous tend les indicateurs. Enfin, aux sections 9 et 10, nous présentons sur des données réelles provenant d'une grande enquête réalisée par Statistics Sweden. Le deuxième (section 10) décrit une simulation exécutée sur une population finie synthétique.

## 2. Estimateurs par calage pour une enquête avec non-réponse

Un échantillon probabiliste  $s$  est tiré de la population  $U = \{1, 2, \dots, k, \dots, N\}$ . Le plan d'échantillonnage donne à l'unité  $k$  la probabilité d'inclusion connue  $\pi_k = \Pr(k \in s) > 0$  et le poids de sondage connu  $d_k^k = 1/\pi_k$ . Des cas de non-réponse ont lieu. L'ensemble de réponses  $r$  est un sous-ensemble de  $s$ ; la façon dont il a été généré est inconnue. Nous supposons que  $r \subset s \subset U$ , et que  $r$  est un ensemble non vide. Le taux de réponse (pondéré par les poids de sondage) est

$$p = \frac{\sum_r d_k^k}{\sum_U d_k^k} \quad (2.1)$$

(si  $A$  est un ensemble d'unités,  $A \subseteq U$ , la somme  $\sum_{k \in A}$  s'écrit  $\sum_A$ ). Habituellement, de nombreuses variables sont étudiées dans le cadre d'une enquête. Une variable étudiée typique, qu'elle soit continue ou catégorique, est désignée par  $y$ . Sa valeur pour l'unité  $k$  est  $y_k$ , enregistrée pour  $k \in r$ , et non disponible pour  $k \in U - r$ . Nous cherchons à estimer le total  $Y$  de population,  $Y = \sum_U y_k$ . De nombreux paramètres d'intérêt dans la population finie sont des fonctions de plusieurs totaux, mais nous pouvons nous concentrer sur un seul d'entre eux.

L'information auxiliaire est de deux types, auxquels correspondent deux types de vecteurs,  $\mathbf{x}_*^k$  et  $\mathbf{x}_0^k$ . L'information auxiliaire de population est transmise par  $\mathbf{x}_*^k$ , une valeur vectorielle connue pour chaque unité  $k \in U$ . Donc,  $\sum_U \mathbf{x}_*^k$  est un total de population connu. Alternativement, nous permettons que  $\sum_U \mathbf{x}_*^k$  soit importé d'une source extérieure et que  $\mathbf{x}_0^k$  soit une valeur vectorielle connue (observée) pour chaque unité  $k \in s$ . L'information

# Plan d'estimation : détermination de vecteurs auxiliaires en vue de réduire le biais de non-réponse

Carl-Erik Särndal et Sixten Lundström<sup>1</sup>

## Résumé

Le présent article décrit l'élaboration d'outils de calcul, appelés indicateurs, qui permettent de juger de l'efficacité de l'information auxiliaire utilisée pour contrôler le biais de non-réponse dans les estimations par sondage, obtenues ici par calage. L'étude est motivée par le contexte dans lequel sont réalisées les sondages dans plusieurs pays, surtout en Europe du Nord, où de nombreuses variables auxiliaires possibles concernant les ménages et les particuliers sont tirées de registres administratifs fiables. Un grand nombre de vecteurs auxiliaires pouvant donc être composés, il est nécessaire de les comparer afin de déterminer dans quelle mesure ils peuvent réduire le biais. Les indicateurs décrits dans le présent article sont conçus pour répondre à ce besoin. Ils sont utilisés dans les enquêtes réalisées par Statistics Sweden. Nous considérons des conditions générales d'enquête où un échantillon probabiliste est tiré de la population finie selon un plan d'échantillonnage arbitraire et où des cas de non-réponse se produisent. La probabilité d'inclusion dans l'échantillon est connue pour chaque unité de la population ; la probabilité de réponse est inconnue, ce qui cause un biais. La variable étudiée (variable  $y$ ) n'est observée que pour l'ensemble de répondants. Quel que soit le vecteur auxiliaire utilisé dans un estimateur par calage (ou dans toute autre méthode d'estimation), un biais résiduel persiste systématiquement. Le choix du vecteur auxiliaire (le meilleur possible) est guidé par les indicateurs proposés dans le présent article. Dans les premières sections, nous décrivons le contexte de leur élaboration et leurs caractéristiques de calcul, puis nous exposons leur contexte théorique. Les dernières sections sont consacrées aux études empiriques. L'une de ces études illustre la sélection des variables auxiliaires dans une enquête réalisée par Statistics Sweden. Une deuxième illustration empirique consiste en une simulation à partir d'une population finie synthétique ; un certain nombre de vecteurs auxiliaires possibles sont classés par ordre de préférence à l'aide des indicateurs.

Mots clés : Pondération par calage ; correction de la non-réponse ; biais de non-réponse ; variables auxiliaires ; indicateur de biais.

## 1. Introduction

De nos jours, on peut s'attendre à un taux de non-réponse élevé dans de nombreuses enquêtes, de sorte qu'il faut élaborer des méthodes qui permettent de réduire autant que possible le biais de non-réponse dans les estimations. Il faut donc disposer d'information auxiliaire puissante. Les fichiers de données administratives sont l'une des sources de ce genre d'information. À cet égard, les pays scandinaves et certains autres pays européens, notamment les Pays Bas, sont dans une situation avantageuse. De nombreuses variables auxiliaires possibles (appelées variables  $x$ ) peuvent être tirées de registres administratifs de haute qualité dans lesquels les valeurs des variables auxiliaires sont spécifiées pour l'entière de la population. Les variables mesurant divers aspects de la collecte des données représentent une autre catégorie utile de données auxiliaires. Des mesures efficaces peuvent être prises pour contrôler le biais de non-réponse. Au delà du plan d'échantillonnage, le *plan d'estimation* devient, dans ces pays, une composante importante du plan d'enquête global. Statistics Sweden a consacré des ressources considérables à l'élaboration de méthodes en vue de sélectionner les meilleures variables auxiliaires possibles.

De nombreux articles traitent de la pondération dans les sondages présentant une non-réponse et de la sélection des « meilleures variables auxiliaires ». Eltinge et Yansaneh (1997), Kalton et Flores-Cervantes (2003), ainsi que Thomsen, Wang et Zhang (2006) en sont des exemples. Une attention particulière est accordée à la pondération dans le cas des enquêtes par panel avec érosion du panel dans, par exemple, Rizzo, Kalton et Brick (1996), selon lesquels « le choix des variables auxiliaires est important, probablement plus important que celui de la méthode de pondération ». La revue effectuée par Kalton et Flores-Cervantes (2003) fournit de nombreuses références à des travaux antérieurs. Comme dans le présent article, Deville (2002) et Kott (2006) privilégient une approche de pondération par calage pour corriger la non-réponse. Certaines méthodes antérieures sont des cas particuliers de la perspective adoptée dans le présent article, laquelle est fondée sur l'utilisation systématique d'information auxiliaire en effectuant un calage à deux niveaux. Récemment, la recherche d'une méthode de pondération efficace a pris deux directions, à savoir i) fournir des conditions plus générales que les méthodes populaires, mais limitées, de pondération par cellule et ii) quantifier la recherche de variables auxiliaires à l'aide d'indicateurs calculables.





plein. Le reste du budget est affecté au financement de travaux de recherche à temps partiel effectués par quelque 70 autres méthodologistes. Ces dispositions particulières ont plusieurs objectifs. Premièrement, elles contribuent à la pertinence de la recherche. Deuxièmement, elles contribuent à l'adoption des résultats de la recherche. Enfin, troisième-ment, elles soutiennent le moral, car, si tout le monde ne désire pas faire de la recherche (ou n'est pas capable d'en faire), nombreux sont ceux qui souhaitent s'y essayer et le simple fait d'effectuer certains travaux de recherche, chez ceux qui en sont capables, ouvre l'esprit et se traduit par une pratique plus éclairée.

Nous veillons autant que possible à ce que les projets de recherche particuliers qui sont approuvés soient en alignement avec les priorités de recherche générales de Statistique Canada, tout en laissant une certaine latitude pour le lancement personnel de travaux de recherche. Pour cela, nous établissons chaque année les grandes priorités et invitons les membres du personnel à soumettre des propositions dans ces domaines. Les propositions font l'objet d'un processus de décision officiel en vue de choisir les meilleures et d'y affecter des employés. Le directeur de la Division de la recherche et de l'innovation en statistique et son petit effectif permanent offrent des conseils et un encadrement.

Les mesures supplémentaires qui suivent visent à assurer la qualité des travaux de recherche effectués :

- La possibilité de publier des articles dans *Techniques d'enquête*, qui est la propre revue de Statistique Canada, joue un rôle d'incitatif. Alors que l'examen des articles par les pairs est géré rigoureusement par un comité de rédaction international, l'existence d'un débouché local, mais prestigieux, pour la recherche en méthodologie est un témoin visible de l'engagement de la haute direction.
- Nous participons régulièrement à la rédaction d'articles avec d'autres chercheurs externes réputés (Canadiens et non-Canadiens).
- Nous participons régulièrement à des échanges méthodologiques avec les méthodologistes du Bureau of the Census et du Bureau of Labour Statistics des États-Unis.
- Nous participons activement aux activités d'organismes statistiques canadiens, américains et internationaux.

## 10. Conclusion

Comme il est mentionné dans l'introduction, l'article est consacré en grande partie aux outils qui devraient être pris en considération par les bureaux de la statistique en vue d'établir et de soutenir la fonction de méthodologie et la recherche connexe, outils qui, en combinaison appropriée,

## Bibliographie

peuvent accroître à la fois l'indépendance professionnelle et la pertinence de la fonction. Je tiens toutefois à souligner qu'il ne s'agit pas d'un livre de recettes. L'élément qui importe plus que tous les outils est le milieu environnant : le bureau de la statistique accepte-t-il volontiers les questions et veille-t-il à ce que l'on réponde de façon substantielle aux questions de fond ? Le changement est-il intrinsèquement mal accepté ? La collégialité est-elle favorisée ? La prise de risque intelligente est-elle encouragée ou désapprouvée ? Les expériences sont-elles les bienvenues, évaluées en fonction de leur mérite et suivies de prises de décision ? Il s'agit là d'attributs impartis par la haute direction du bureau de la statistique et les outils ne peuvent pas les remplacer. S'il n'a pas le leadership qu'il lui faut, le meilleur personnel de méthodologie (ou en fait, le meilleur bureau de la statistique) déperdra. Mais l'inverse n'est pas vrai : il est essentiel de bien comprendre les équilibres subtils recommandés dans le présent article, et de déployer prudemment les outils qui permettent de les réaliser. Mais même dans ces conditions, seule une stratégie de long terme peut être couronnée de succès.

Je suis entièrement convaincu que Joe serait d'accord avec ma conclusion (Waksberg 1998).

Brackstone, G.J. (1991). Shaping statistical services to satisfy user needs. *Statistical Journal of the United Nations Economic Commission for Europe*, 8, 3/4, 243-258.

Brackstone, G. (1997). Organization of a survey methodology service. Dans *Enquêtes et sondages : méthodes, modèles, applications, nouvelles approches*, (Eds., G. Brossier et A.-M. Dussaix), Rennes, France, 19-20 juin, 3, 118-134.

Fellegi, I.P. (1991). Maintaining public confidence in official statistics. *Journal of the Royal Statistical Society, Series A* (Statistics in Society), 154, Part 1, 1-6.

Fellegi, I.P. (1992). Planning and Priority Setting - the Canadian Experience. Dans *Statistics in the Democratic Process at the End of the 20<sup>th</sup> Century*, (Eds., Hölder, Malaguerria et Vukovich), Anniversary publication for the 40<sup>th</sup> Plenary Session of the Conference of European Statisticians. Publié par le Federal Statistical Office, Wiesbaden, Federal Republic of Germany.

Fellegi, I.P. (1996). Characteristics of an effective statistical system. *Revue Internationale de Statistique*, 64, 2, 165-199.

Fellegi, I.P. (2004). Maintaining the credibility of official statistics. *Statistical Journal of the United Nations*, ECE 21, 191-198.

Statistique Canada (1995). Formation et perfectionnement à Statistique Canada. Institut de formation de Statistique Canada, mars 1995.

Waksberg, J. (1998). The Hansen Era: Statistical research and its implementation at the U.S. Census Bureau, 1940-1970. *Journal of Official Statistics*, 14, 2, 119-135.

- fait l'objet d'un bref exposé oral donné par des employés.
- Pour chaque point à l'ordre du jour, un membre du Comité est désigné comme commentateur officiel. Les commentateurs présentent formellement leur point de vue. Étant donné que la plupart des documents de référence sont préparés par des employés à mi-chemin dans leur carrière, ces discussions sont non seulement très fructueuses pour les projets qui en sont l'objet, mais elles contribuent aussi à la formation des employés concernés ainsi qu'à celle des autres participants.
- Les réunions du Comité comptent la présence d'un grand nombre de méthodologistes, mais aussi des membres de la direction de la division spécialisée concernée, y compris, souvent, le statisticien en chef et un ou deux de ses adjoints.
- Le Comité se réunit régulièrement : deux fois par an, pendant un jour et demi à chaque occasion.
- Les membres du Comité examinent régulièrement le suivi donné à leurs conclusions et recommandations officielles, ce qui leur permet de vérifier si leurs conseils sont pris au sérieux.

## 9. Recherche

### Réflexions générales

Je tiens pour acquis qu'il ne soit pas nécessaire ici de passer du temps à souligner l'importance intrinsèque de la recherche dans un bureau de la statistique. J'insisterai toutefois sur les points suivants :

- L'organisation de la fonction de recherche doit faire l'objet d'une mûre réflexion de manière à maximiser à la fois sa pertinence et la probabilité que ses résultats seront intégrés avec succès dans la pratique quotidienne. Deux dangers doivent être évités : que la recherche soit inspirée par l'intérêt personnel ou qu'elle soit tellement axée sur la tâche qu'elle en soit dépourvue d'imagination ;
- la recherche doit être financée adéquatement ;
- le programme de recherche interne doit établir et entretenir des liens étroits avec les programmes de recherche extra-muros pertinents.

### Le cas de Statistique Canada

L'une des quatre divisions de la méthodologie se consacre officiellement à la recherche à temps plein. Cependant, l'organisation des travaux de recherche est particulière. Même si le budget de recherche prévoit l'équivalent de 22 chercheurs à temps plein, la division de la recherche proprement dite n'est dotée que de 6 membres à temps

## 8. Comité consultatif

### Réflexions générales

Le perfectionnement professionnel comporte bien plus que la formation. Les employés, surtout au début de leur carrière, ont régulièrement l'occasion de travailler sur divers types de projets, comme des enquêtes démographiques, socioéconomiques et autres des entreprises, l'utilisation de données administratives ou le couplage d'enregistrements. Nombre d'entre eux assistent aussi à des conférences scientifiques. Par exemple, ces dernières années, quelque 17 % des employés de la méthodologie par année ont participé à diverses conférences professionnelles au Canada et à l'étranger. Les employés sont également encouragés à travailler sur des projets de recherche et à publier les résultats dans des revues à comité de lecture, y compris la revue *Techniques d'enquête* de Statistique Canada. Enfin, depuis de nombreuses années maintenant, Statistique Canada organise un symposium international sur les questions de méthodologie auquel sont invitées des chefs de file de la recherche venant de partout dans le monde. Ces symposiums sont évidemment ouverts à tous les employés de Statistique Canada et la plupart des méthodologistes choisissent d'y participer.

### Le cas de Statistique Canada

Un comité consultatif de la méthodologie peut remplir une fonction des plus utile, à savoir a) veiller à ce que soient adoptées de bonnes pratiques méthodologiques, b) intégrer ces pratiques dans les activités quotidiennes des organismes statistiques et c) assurer la formation du personnel. Toutefois, le comité ne peut être efficace que si : a) son avis est sollicité au sujet de questions de méthodologie importantes et b) des mécanismes sont mis en place pour s'assurer que soit accordé aux avis du Comité le poids qui leur est dû. J'ai observé des comités consultatifs de la méthodologie jouant un rôle tout aussi utile dans un bureau centralisé (Statistique Canada) que dans un bureau décentralisé (le Bureau of the Census au cours des années 1960).

Le Comité consultatif de la méthodologie de Statistique Canada joue un rôle clé. Plusieurs facteurs contribuent à son utilité et à son autorité :

- La réputation personnelle des membres du Comité est l'un des éléments.
- Chaque projet important de Statistique Canada est présenté au Comité afin d'obtenir son avis.
- Pour faciliter l'examen du Comité, pour chaque point à l'ordre du jour, un document de référence est préparé et



techniques à prendre en considération dans la recherche de l'équilibre optimal ;

- sur le plan tactique, les méthodologues fournissent les méthodes et les outils statistiques qui sont intégrés dans la conception globale de l'enquête, c'est-à-dire le plan de sondage, les méthodes d'estimation et de pondération, le contrôle de la qualité, les stratégies de vérification et d'imputation, les vérifications de la couverture, les méthodes d'analyse et ainsi de suite.

Les équipes de projet fonctionnent le mieux dans une structure organisationnelle où les décisions sont axées sur le mérite, où tout le monde peut poser des questions et s'attendre à des réponses réfléchies, c'est-à-dire une organisation conçue pour maximiser l'exploitation des compétences de toutes les personnes concernées.

## 7. Perfectionnement professionnel

### *Réflexions générales*

Indispensable pour tous les groupes professionnels, le perfectionnement professionnel englobe la formation officielle, ainsi que des approches formelles et non formelles en vue de faciliter l'apprentissage en cours d'emploi. À mon avis, les employés de la méthodologie requièrent à cet égard une attention particulière ; parce qu'en général, les universités n'offrent que peu, voire aucun, cours de méthodologie d'enquête. (Il existe de plus en plus d'exceptions, mais leur nombre reste loin d'être écrasant. L'une des plus notables est le Joint Program in Survey Methodology de l'Université du Maryland. Mais il existe également au Royaume-Uni, en Irlande et en Nouvelle-Zélande des programmes menant à un grade en statistique officielle qui englobent la méthodologie d'enquête.) Puisque des connaissances professionnelles approfondies sont essentielles pour satisfaire aux exigences de pertinence et d'indépendance, la plupart des bureaux de la statistique qui désirent maintenir un effectif de méthodologistes solide n'ont pas d'autre choix que de se doter d'un programme de perfectionnement professionnel conçu minutieusement, que la fonction de méthodologie soit centralisée ou décentralisée.

Àfin que les cours soient pertinents, il est souhaitable qu'une proportion importante de ceux-ci soit donnée par des membres du personnel qui sont eux-mêmes des praticiens actifs. Cette approche est plus facile à mettre en place dans les organisations centralisées où les méthodologistes principaux peuvent non seulement affecter certains employés à des tâches d'enseignement (habituellement à temps partiel), mais peuvent aussi prendre des dispositions pour les remplacer dans les travaux des projets courants.

Les aspects plus généraux du perfectionnement professionnel sont également plus faciles à prendre en considération dans les organisations centralisées : celles-ci peuvent plus aisément gérer les affectations périodiques des employés à divers types de travaux d'enquête, la participation à des conférences scientifiques, l'offre de possibilités de recherche à ceux cela intéresse et qui sont capables d'effectuer des travaux de recherche à temps partiel et, fait le plus important, le travail de stagiaires sous la supervision de méthodologistes plus expérimentés.

### *Le cas de Statistique Canada*

Statistique Canada met l'accent sur la formation en général, sans se limiter à la méthodologie (voir Statistique Canada 1995). Globalement, les dépenses en formation du Bureau représentent environ 3 % de son budget (15 millions de dollars), lesquels sont consacrés à la formation officielle – à cela s'ajoutent beaucoup d'autres dépenses consacrées à divers moyens de perfectionnement. En revanche, étant donné le rôle central de la formation en méthodologie, le pourcentage du budget de la méthodologie consacré à cette formation est près de deux fois plus élevé (s'approchant des 6 % pour l'exercice 2008-2009).

La formation est prodiguée sous forme de cours officiels donnés à l'Institut de formation de Statistique Canada qui, à l'heure actuelle (en 2009), offre quelque 20 cours de méthodologie, dont le niveau varie des cours d'introduction à des matières de niveau universitaire supérieur. La plupart des cours sont donnés par des employés de Statistique Canada ; à l'occasion, des universitaires, provenant principalement des universités locales, sont engagés si cela les intéresse de prodiguer un enseignement et (ou) de contribuer au perfectionnement de nos employés d'autres façons (par exemple consultation). Dans ce dernier cas, nous avons été particulièrement fortunés de pouvoir compter sur la contribution du professeur J.N.K. Rao, de l'Université Carleton, pendant plusieurs décennies.

Tous les nouveaux employés doivent suivre un cours de base de six semaines (avec travaux pratiques à l'appui) portant sur le poids de sondage, les opérations d'enquête, ainsi que le traitement et l'analyse des données. Les objectifs de ce cours d'introduction sont multiples. Puisque le même cours de base de six semaines concernant le travail d'enquête est suivi par *tous* les nouveaux professionnels, il permet d'inculquer très tôt à chacun les connaissances élémentaires sur tout ce que comporte le travail d'enquête, et, fait encore plus important, de leur faire saisir l'importance critique du travail en équipe interdisciplinaire. C'est également à ce stade que les nouveaux employés provenant d'autres disciplines prennent connaissance pour la première fois des exigences méthodologiques concernant la conception des enquêtes.



suffisamment de poids aux avis réfléchis des méthodologistes.

Dans les organisations décentralisées, les équipes de projet assurent de façon tout aussi importante pour que les opinions des méthodologistes soient prises en considération comme il se doit. Mais ici, les chances jouent clairement en faveur de la pertinence au détriment de l'indépendance. Or, accorder trop de poids à la « pertinence » a également ses risques, car cela peut donner lieu à une optimisation locale.

Cette dernière est une situation où les enquêtes sont optimisées sans tenir compte des objectifs généraux du bureau de la statistique. Par exemple, les enquêtes pour-raient être personnalisées au point qu'il devienne difficile d'introduire des gains d'efficacité importants grâce à l'utilisation de normes collectives et de systèmes généraux (l'usage à grande échelle d'approches, de systèmes et d'outils généralisés peut être une source d'économies et considérables à l'échelle du bureau, parce que les temps de mise en œuvre sont raccourcis, les dépenses en développe-ment et en maintenance des systèmes sont réduites, la rotation du personnel est facilitée, etc.). Toutefois, les systèmes généralisés pourraient être dépourvus de certaines fonctionnalités pouvant accroître l'efficacité d'une opération donnée. Les organisations décentralisées sont plus sus-ceptibles que celles qui ne le sont pas de privilégier ce genre de solutions élaborées localement plutôt que les outils standard intégrés, même si ces derniers peuvent donner lieu dans le long terme à des économies importantes.)

*Centralisation : le cas de Statistique Canada*

À Statistique Canada, les équipes affectées à d'im-portants projets de développement doivent rendre des comptes à des comités directeurs habituellement composés des chefs des disciplines participantes. Le comité directeur approuve la stratégie générale du projet et sert, au besoin, de forum pour l'examen des questions que les équipes ne sont pas capables de résoudre elles-mêmes. En pratique, les appels de ce genre sont rares et limités aux cas concernant des principes professionnels ou des questions vraiment stratégiques. Le comité directeur veille à ce que les problèmes soient résolus au sein de l'équipe de projet non pas en fonction du rang, mais plutôt en fonction du mérite professionnel.

La fonction des méthodologistes qui font partie des équipes de projet est double :

• Sur le plan stratégique, ils veillent à ce que la conception générale de l'enquête réponde aux objectifs de fond du projet, tout en établissant un juste équilibre entre la fiabilité, le coût, l'actualité et le fardeau de réponse. Bien que l'équipe de projet au complet soit concernée par l'établissement de cet équilibre, ce sont les méthodologistes qui fournissent le cadre et les

3. Une troisième composante provient des ressources supplémentaires financées directement par les divisions spécialisées bénéficiaires qui, en fait, économisent sur leurs autres dépenses pour pou-voir utiliser des services de méthodologie addition-nels. Ces fonds supplémentaires représentent une part non négligeable, de quelque 20 %, du budget de la méthodologie. Le fait même que les divisions spécialisées considèrent la fonction de méthodo-logie suffisamment valable pour financer directe-ment l'obtention de conseils méthodologiques en dit long quant à la santé de la relation et la mesure dans laquelle les services sont appréciés. Les fonds en question sont consacrés à divers projets incluant des améliorations, sauf le remaniement important de projets en cours. Ce financement sensibilise aussi davantage les méthodologistes à l'idée que leurs activités doivent demeurer pertinentes pour leurs utilisateurs. Le genre de services qu'ils fournissent a une influence directe sur la quantité des ressources qui sont mises à leur disposition.

4. La quatrième composante du budget de la métho-dologie (environ 20 %) correspond aux projets financés extérieurement, qui sont habituellement des enquêtes financées par d'autres ministères. Elle ne nécessite pas d'autres commentaires

5. La dernière composante (7 %) est consacrée à la recherche. Il s'agit d'un « financement global », ce qui signifie qu'un montant fixe de financement est affecté à la fonction de recherche. L'affectation annuelle est régie par un mécanisme qui sera décrit plus loin.

La complexité du mécanisme de financement et la multiplicité des sources de financement témoignent du soin avec lequel le Bureau cherche à équilibrer les vertus de l'indépendance et celles de la pertinence.

## 6. Équipes de projet

### *Réflexions générales*

La création d'équipes pour exécuter les projets de déve-

loppement renforce la pertinence sans que cela nuise à l'indépendance. Cependant, les équipes de projet ne sont pas une panacée, car tout dépend de la mise en place de freins et contrepois appropriés. Dans les organisations centralisées, les équipes de projet, le plus souvent dirigées par un gestion-naire de projet issu du secteur spécialisé commanditaire, permettent d'amener plus facilement les méthodologistes participants à accorder l'attention qu'il convient aux objectifs et contraintes des projets. Persiste néanmoins le risque intrinsèque que le gestionnaire de projet ne donne pas

disciplines clés) un forum explicite pour faire part de leur opinion au stade embryonnaire critique des nouveaux projets.

*Centralisation : le modèle de Statistique Canada*

Chaque nouveau projet ou remaniement important est approuvé dans le cadre du système de planification de Statistique Canada. En prévision de son examen, un budget exhaustif est établi et chacune des principales disciplines qui participeront au projet approuve la conception et les modalités opérationnelles proposées si elle les juge appropriées. Si le projet est approuvé, son budget est réparti entre les disciplines participantes, y compris la méthodologie. À leur tour, ces organisations « s'engagent » à livrer les produits et services convenus en respectant les budgets approuvés. Un gestionnaire de projet supervise l'avancement des travaux et les dépenses, et est autorisé à réaffecter les ressources, au besoin.

Le budget de la fonction de méthodologie comprend cinq sources distinctes destinées, d'une part, à faciliter la bonne planification de l'utilisation de la méthodologie et son intégration complète dans le travail du Bureau et, d'autre part, à obtenir le financement nécessaire.

1. La participation de la méthodologie aux projets *développés* est garantie par le processus de planification de Statistique Canada comme je l'ai mentionné plus haut. La contribution financière de ces projets au budget de la méthodologie peut varier d'année en année, mais la stabilité globale est raisonnable (ce qui facilite le recrutement et le perfectionnement du personnel permanent). Ces projets représentent près de 30 % du budget total de la méthodologie. Ils comportent habituellement des remaniements importants, qui nécessitent souvent une expérimentation et une innovation considérable.

2. Mais l'intervention de la méthodologie est également importante. L'ajustement mineur de la conception, *etc.*, Des compris l'estimation de la variance s'il y a lieu, la qualité, surveillance des diverses erreurs, y compris l'ajustement mineur de la conception, *etc.*, Des ressources de base sont réservées à ces activités et réparties de manière plus ou moins permanente entre les grands secteurs spécialisés. Ces ressources, qui sont la deuxième composante du budget de la méthodologie, représentent un peu moins de 25 % de celui-ci. Alors que, pour la méthodologie, les activités « courantes » représentent moins de 25 % de la charge de travail, pour Statistique Canada dans son ensemble, les travaux « courants » représentent plus de 90 % du budget. Cette différence reflète le caractère novateur des travaux de méthodologie.

autorisée des ressources dont il a la responsabilité aux projets les plus stratégiques. À Statistique Canada, le principal défenseur d'une bonne méthodologie a le statut de statisticien en chef adjoint (SCA), qui est le rang directement inférieur à celui de statisticien en chef du Canada. Afin d'obtenir un poste aussi élevé dans une bureaucratie gouvernementale, la responsabilité hiérarchique du SCA (méthodologie) englobe les normes statistiques (classification et registres centraux), ainsi que l'informatique (TI). Bien que le poste comporte des responsabilités qui s'étendent au-delà de la méthodologie, il est de tradition depuis longtemps (plus de 35 ans) qu'il soit occupé par un expert reconnu en méthodologie, qui peut donc faire part avec autorité à la haute direction de l'importance de cette dernière en général, ainsi que dans le contexte de projets particuliers.

## 5. Planification et financement

*Réflexions générales*

Le bon fonctionnement de la méthodologie (et, en fait, de tout le bureau de la statistique) dépend fortement de l'existence d'un système de planification approprié (voir Fellegi 1992 et Brackstone 1991).

- La planification est une condition nécessaire à la répartition rationnelle des ressources en tout temps.
- Elle sert également à marquer explicitement le début et la fin des projets de développement et, par conséquent, représente l'occasion idéale pour la méthodologie d'« approuver » la conception proposée des nouveaux projets.
- Enfin, le système de planification offre à la méthodologie l'occasion de juger explicitement si un nouveau projet planifié pourra être réalisé en respectant simultanément les contraintes budgétaires, les normes de qualité du bureau et les coûts prévus d'entretien. En fait, le système de planification donne aussi aux représentants de toutes les disciplines qui participent à la création d'un nouveau projet (sa planification ou sa mise en œuvre) la possibilité de « donner leur approbation » afin de signaler qu'ils assument la responsabilité professionnelle de l'adéquation de son financement ou de l'intégrité de son fonctionnement.

Un tel système de planification est essentiel quand les principales disciplines (méthodologie, développement des systèmes, collecte des données, *etc.*) sont centralisées, car autrement, les organisations responsables ne peuvent pas prendre des dispositions pour obtenir les ressources nécessaires. Cependant, pour des raisons plus subtiles, les bureaux décentralisés en ont tout autant besoin, afin de donner aux leaders de la méthodologie (et, en fait, d'autres



lesquels ils conçoivent des enquêtes. Cette mesure supplémentaire est prise afin de s'assurer qu'ils se concentrent sur les bonnes questions.

5. Enfin, au nom des pratiques exemplaires, les méthodologistes effectuent des sondages sur la satisfaction des clients qui fournissent des renseignements sur tous les aspects de leur travail et, avant tout et par-dessus tout, sur sa pertinence, et prennent des mesures de suivi fondées sur les résultats.

#### 4. Leadership

##### *Réflexions générales*

Le leadership est essentiel. En plus d'une formation

universitaire appropriée et d'une grande expérience de la méthodologie, le leader de la fonction de méthodologie doit avoir une vision stratégique et une personnalité qui inspire confiance. Il s'agit d'une fonction intrinsèquement difficile. Dans l'écrasante majorité des bureaux de la statistique, les considérations opérationnelles et celles ayant trait aux domaines spécialisés sont celles qui suscitent le plus d'attention. Dans ce genre d'environnement, une voix faisant autorité est nécessaire afin de s'assurer que soient obtenues des ressources adéquates pour la fonction de méthodologie proprement dite, mais avant tout, en vue d'orienter le *bureau dans son ensemble* dans des directions techniquement valables et, inversement, afin de mettre un frein aux initiatives qui ne peuvent pas bénéficier d'un fondement méthodologique solide. Un « fondement solide » signifie davantage qu'une bonne conception des enquêtes s'appuyant sur les meilleures connaissances courantes disponibles. Cela comprend aussi la notion de planification stratégique de la recherche, des expériences et des enquêtes pilotes, en vue d'accroître la probabilité que tout le savoir qui sera nécessaire dans l'avenir sera disponible. Pour avoir de l'effet, les opinions des méthodologistes doivent être appuyées par un leader qui possède non seulement une compétence personnelle incontestée, mais aussi une place parmi la haute direction de l'organisme statistique.

Si les méthodologistes n'appartiennent pas à une organisation centrale au sein du bureau de la statistique, il est d'autant plus important que la position hiérarchique de leur représentant supérieur soit élevée, puisque sous un scénario décentralisé, il ne possèdera aucune autorité directe pour ce qui est (de la plupart) des ressources méthodologiques.

*Centralisation : le modèle de Statistique Canada*

La centralisation offre au leader de la fonction de méthodologie un autre levier à jouer son rôle, car elle lui permet de procéder à l'affectation rationnelle et

hiérarchie dépendant de l'effectif qu'il contrôle directement, sans qu'aucune disposition – comme il en existe dans certains pays – ne soit prise pour que son niveau d'accès et sa position sur l'échelle hiérarchique dépendent de son prestige personnel plutôt que de la taille ou du niveau du personnel de soutien.

##### *Centralisation : le modèle de Statistique Canada*

Il y a de nombreuses années, Statistique Canada a opté pour le modèle centralisé (voir Fellegi 1996) et ce choix n'a jamais été sérieusement mis en doute (il l'a été brièvement à la fin des années 70, mais, concrètement, la mise en question n'a mené nulle part); Statistique Canada a également établi un certain nombre de pratiques destinées à réduire la menace tenant au fait que la centralisation pourrait réduire la pertinence.

1. Équipes de projet : Ces équipes sont interdisciplinaires et comprennent d'office un méthodologiste, mais sont dirigées par un gestionnaire de projet dont l'association à ce dernier relève du domaine spécialisé et qui assume vraisemblablement la responsabilité opérationnelle de l'entité de projet.

2. Financement : le financement de la fonction de méthodologie est contrôlé en grande partie par le reste de Statistique Canada. Les secteurs de programme (dont je décrirai les limites plus loin) sont libres de consacrer ou non leur agent à l'achat de services de méthodologie à condition qu'il se conforme aux normes de qualité et aux normes reconnues du Bureau. Étant donné que leur budget est en grande partie remis en question année après année, cette responsabilisation signifie que les méthodologistes ont tout intérêt à être sensibles aux besoins des programmes du Bureau.

3. Organisation de la fonction de méthodologie : elle ressemble en grande partie à celle de Statistique Canada. Le Bureau compte quatre divisions de la méthodologie, dont trois fournissent des services de méthodologie à trois secteurs différents du Bureau, tandis que la quatrième se consacre à la recherche. En fait, les trois divisions de la méthodologie appliquée sont elles-mêmes organisées par domaine spécialisé en suivant le modèle d'organisation du Bureau (la rotation régulière des employés de la méthodologie fait en sorte que des occasions de perfectionnement général sont offertes aux méthodologistes).

4. Regroupement des employés de la méthodologie dans les secteurs spécialisés : les méthodologistes sont de temps en temps transférés physiquement dans les locaux des secteurs spécialisés pour



« centralisation » et « décentralisation », car quelle que soit la modalité organisationnelle fondamentale adoptée, de nombreux outils supplémentaires sont nécessaires pour compenser ses inconvénients tout en conservant ses avantages intrinsèques. La suite de l'article s'articule donc sur une discussion des principaux outils (en choisissant ces outils comme sujet de discussion, je me suis inspiré de l'article de Brackstone 1997) repris sous les thèmes suivants :

- organisation ;
- leadership ;
- planification et financement ;
- équipes de projets ;
- perfectionnement professionnel ;
- comités consultatifs ;
- interaction avec le monde universitaire ;
- recherche.

### 3. Organisation

#### Réflexions générales

Les bureaux nationaux de la statistique se distinguent par la façon dont ils organisent leur fonction de méthodologie. Certains la répartissent entre des éléments individuels de l'organisme, chacun responsable d'un sujet donné (par exemple le travail). D'autres procèdent à une décentralisation partielle, par exemple entre des domaines spécialisés plus généraux (tels que la démographie ou la statistique des entreprises). Ainsi, le Bureau of the Census des États-Unis a décentralisé en grande partie sa fonction de méthodologie. En revanche, Statistique Canada et l'Australian Bureau of Statistics l'ont en grande partie centralisée. De nombreux facteurs influencent le choix organisationnel. Par exemple, en France et en Inde, où tous les professionnels ont la même formation statistique et sont pour la plupart recrutés auprès d'un seul établissement d'enseignement, l'accent est évidemment mis sur la centralisation de la formation et dans une moindre mesure, de la recherche.

Les arguments classiques sont que la décentralisation favorise la pertinence, tandis que la centralisation favorise l'indépendance. Cependant, le but devrait être d'obtenir les deux. D'où la question de savoir comment nous pouvons accroître l'indépendance, dans le cas d'une organisation décentralisée de la méthodologie, et la pertinence, dans le cas d'une organisation centralisée.

Bien qu'elle puisse servir à souligner la pertinence, la décentralisation possède des inconvénients intrinsèques. Puisque les unités entre lesquelles la fonction de méthodologie est décentralisée sont nécessairement plus petites qu'elles ne le seraient dans des conditions de plus grande centralisation, elles sont moins susceptibles de faciliter la

spécialisation et la recherche. Elles sont aussi moins susceptibles de favoriser un échange d'idées avec les méthodologistes se penchant sur d'autres questions. En outre, puisque les fonctions d'exécution entre lesquelles la méthodologie est décentralisée ne sont habituellement pas dirigées par des méthodologistes, ce modèle a tendance à donner une position hiérarchique moins élevée aux chefs des unités de méthodologie décentralisées. En cas de « conflits » – et ceux-ci sont inévitables étant donné les perceptions différentes des priorités, des coûts, de la qualité et ainsi de suite – toutes choses étant égales par ailleurs, les méthodologistes auront plus de difficulté à défendre leur opinion professionnelle. Sans contrepois, ce genre d'organisation peut devenir déséquilibré.

Un contrepois essentiel pourrait être un « méthodologiste en chef » qui relève directement du chef du bureau de la statistique et est inévitablement appelé à jouer un rôle important dans la planification de long terme et la répartition des ressources. Le « méthodologiste en chef » pourrait voir son pouvoir renforcé s'il obtient la responsabilité directe d'une fonction de recherche et de développement forte, pouvant servir de « base intellectuelle » aux employés de la fonction décentralisée de méthodologie.

Les équipes de projet, formées pour réaliser les grands travaux de développement, sont un autre moyen important d'accroître l'indépendance dans le cas des organisations centralisées. Les projets de ce genre – nécessairement multidisciplinaires s'ils sont d'importance – sont exécutés par des équipes de projet ponctuelles qui fonctionnent en dehors de l'organisation hiérarchique du bureau. Statistique Canada a accordé beaucoup d'attention à l'organisation des équipes de projet et l'a perfectionnée au fil du temps. L'une des caractéristiques de cette organisation est que, quand des désaccords professionnels surviennent au sein d'une équipe et que celle-ci estime que leur résolution requiert une intervention externe, le conflit est porté à l'attention d'un groupe supérieur dont fait partie un membre du personnel du « méthodologiste en chef » (ce qui est automatique si le méthodologiste provient d'un groupe centralisé). C'est ce groupe directeur supérieur qui peut contribuer à protéger l'indépendance.

La fourniture de certains outils additionnels au « méthodologiste en chef » pourrait également être envisagée : il pourrait être autorisé à mettre sur pied un programme de formation méthodologique dynamique et recevoir le financement nécessaire ; il pourrait se voir attribuer un rôle important dans l'affectation et le perfectionnement professionnel du personnel de la méthodologie ; il pourrait être appuyé par un comité consultatif externe énergique. Ces divers éléments tiennent compte du fait que le rôle du « méthodologiste en chef » est particulièrement délicat et pourrait le devenir encore davantage si sa place dans la

fondées sur une méthodologie valable. Une question importante que pose l'organisation de la fonction de méthodologie est celle de savoir comment trouver le juste équilibre entre sa nature intrinsèquement de type service et le besoin d'offrir un encadrement dynamique et efficace. La plus grande partie de l'article traitera de toutes les mesures nécessaires pour s'assurer de réaliser l'objectif de pertinence. Dans le cas de la *recherche* méthodologique, la pertinence signifie que la recherche est non seulement motivée par le travail appliqué, mais aussi qu'elle l'influence.

#### *Indépendance*

La notion d'indépendance de la fonction méthodologique signifie avoir la capacité d'offrir un encadrement méthodologique solide aux projets, quelles que soit l'organisation hiérarchique qui *peut être débattue, mais non ignorée*, et que ce débat est fondé sur des données concluantes et non sur l'autorité. Donc, ma définition de l'indépendance n'est pas que les méthodologistes devraient être capables de « faire ce qui leur plaît » mais plutôt que leur opinion devrait faire autorité.

L'indépendance est souvent mise en contraste avec la pertinence. Puisque cette dernière concerne l'intégration de la méthodologie dans la pratique, on s'efforce fréquemment d'y arriver en implantant des services méthodologiques directement dans la trame des organismes spécialisés. En revanche, on estime qu'en donnant aux méthodologistes leurs propres structures organisationnelles, on accroit l'indépendance. En ce sens, il existe donc une tension entre les deux concepts. Cependant, selon moi, il n'est pas possible d'atteindre la pertinence si l'on ignore l'encadrement méthodologique, de sorte que des mesures appropriées en vue d'assurer l'indépendance sont nécessaires pour que la pertinence se concrétise.

L'indépendance de la *recherche méthodologique* est différente : elle fait généralement référence à un environnement dans lequel les chercheurs ont un avis prédominant dans le choix des sujets de leurs travaux. Manifestement, offrir aux chercheurs ce genre d'environnement crée une tension permanente entre ce choix et la nécessité de demeurer pertinent en tout temps, particulièrement quand il n'est vraiment pas évident dans le court terme de préciser à quoi tient la pertinence. Dans ma discussion des moyens d'atteindre un équilibre entre la pertinence et l'indépendance de la fonction de méthodologie appliquée ainsi que de la recherche méthodologique, je décrirai non seulement les mesures organisationnelles, mais aussi un large éventail d'outils et de modalités qui devraient être envisagés lorsque l'on poursuit cet objectif. J'utiliserais Statistique Canada comme exemple concret. Je souhaite mettre en relief le fait que le problème est nettement plus compliqué que ne l'évoquent les termes

sera le deuxième thème que j'aborderai. Mais pour commencer, je définirai ce que j'entends dans le présent contexte par *méthodologie*, *pertinence* et *indépendance*.

## 2. Quelques définitions

### *Méthodologie*

Le service unique fourni par la *méthodologie* consiste à maximiser la qualité statistique étant donné un budget imposé (ou inversement). Pour cela, les méthodologistes appliquent des pratiques statistiques fondées sur la théorie statistique ou sur l'observation empirique méthodique. Autrement dit, les méthodologistes sont des magiciens des théories statistiques pertinentes, mais aussi de l'« observation empirique méthodique » quand la théorie formelle fait défaut. Par observation empirique méthodique, j'entends des essais planifiés ou le savoir tiré de l'expérience évaluée analytiquement. Donc, j'inclus toutes les connaissances structurées concernant l'utilisation des méthodes et approches qui ont pour objectif de maximiser la qualité en respectant un budget – ou inversement, de minimiser le budget nécessaire pour atteindre un niveau précis de qualité. Cela englobe des éléments tels que le plan de sondage, l'estimation, la vérification des données, l'imputation, l'exploitation de données administratives, le couplage d'enregistrements, la désaisonnalisation, la conception des questionnaires, la mesure de la précision et l'assurance de la qualité des recensements et des enquêtes et l'utilisation de plans expérimentaux.

Les méthodologistes sont avant tout des statisticiens-mathématiciens et leurs travaux portent sur l'aspect appliqué des sujets auxquels ils s'intéressent. Étant donné la nature interdisciplinaire de la statistique officielle, ils interagissent avec les gestionnaires des enquêtes, les spécialistes de la collecte des données, le personnel des TI, les géographes, les sociologues, les économistes, etc.

*Pertinence*

La méthodologie est *pertinente* si les pratiques quotidiennes du bureau de la statistique sont effectivement



# L'organisation de la méthodologie statistique et de la recherche méthodologique dans les bureaux de la statistique

Ivan P. Fellegi<sup>1</sup>

## Résumé

L'article explore et évalue les approches qu'adoptent les bureaux de la statistique pour s'assurer que l'appui méthodologique dont bénéficient leurs activités statistiques soit efficace. La tension qui existe entre les notions d'indépendance et de pertinence est un thème fréquent : en général, les méthodologistes doivent travailler en étroite collaboration avec le reste de l'organisme statistique pour que leurs travaux soient pertinents, mais ils doivent aussi jouer d'un certain degré d'indépendance leur permettant de mettre en question l'utilisation des méthodes existantes et d'en introduire de nouvelles au besoin. Naturellement, il faut aussi établir un programme de recherche efficace qui, d'une part, possède l'indépendance dont a besoin tout programme de recherche et qui, d'autre part, est suffisamment relié aux activités courantes du bureau de la statistique pour que ses travaux soient motivés par ces activités et y soient intégrés en retour. Les thèmes abordés dans l'article sont les divers modes d'organisation, le leadership, la planification et le financement, le rôle des équipes de projet, le perfectionnement professionnel, les comités consultatifs externes, l'interaction avec le monde universitaire et la recherche.

Mots clés : Méthodologie ; statistique officielle ; organisme statistique ; recherche ; pertinence ; indépendance.

## 1. Introduction

C'est pour moi un grand honneur d'accepter un prix portant le nom de Joe Waksberg. Joe était un ami personnel proche, ainsi qu'un bon ami de Statistique Canada.

C'est durant les dernières années qu'il a passées au Bureau of the Census que j'ai fait la connaissance de Joe, quand Morris Hansen m'a demandé de devenir membre de ce qui était alors un comité consultatif en méthodologie des plus impressionnants du Bureau présidé par Bill Cochran. Plus tard, à la fin des années 1970, Statistique Canada, alors aux prises avec de sérieux problèmes d'image et de gestion interne, a demandé à un groupe d'éminents statisticiens d'examiner ce qui n'allait pas. À ma recommandation, Joe a été l'un des trois sages qui ont été sollicités (les autres étant Richard Rugles et le président, Claus Moser). Joe a accepté immédiatement et, de sa manière discrète inimitable, a donné de précieux conseils à Statistique Canada, le message fondamental très utile étant que nous avions de graves problèmes de gestion, mais que notre méthodologie laissait fort peu à désirer.

Il y a quelques années, le Census Bureau m'a fait l'honneur de m'inviter à donner l'un de leurs exposés magistraux annuels des « sages ». Après avoir objecté vivement que je ne me considérerais pas comme un « sage », j'ai fini par accepter leur aimable invitation. Avec sa courtoisie habituelle, Joe a pris le temps d'assister à mon exposé, alors qu'il était octogénaire, mais encore très occupé par ses fonctions de président du conseil d'administration de WESTAT. Nous avons bavardé un bon moment ; c'est la dernière fois que nous nous sommes vus. Quelle carrière ; quelle vie !

Par conséquent, ce n'est pas seulement un honneur professionnel d'accepter le prix Waksberg, mais aussi un plaisir personnel d'être associé à Joe une fois de plus.

En général, m'a-t-on dit, les lauréats du prix Waksberg donnent un aperçu d'un domaine de la méthodologie. Toutefois, bien que, comme vous le savez, j'aie passé la première moitié de ma carrière en tant que méthodologiste, je ne suis plus praticien depuis quelques décennies – quoique je demeure un ardent défenseur (voir Fellegi 2004). Aussi ai-je pensé associer la première moitié de la ma carrière – consacrée à la méthodologie – à la deuxième moitié, consacrée à la gestion des bureaux de la statistique. Je parlerai donc des leçons que j'ai apprises quant à l'organisation du travail de méthodologie appliqué et de la recherche méthodologique dans les bureaux nationaux de la statistique, c'est-à-dire ce qui donne de bons résultats et ce qui est moins fructueux (en supposant que sont satisfaites les conditions essentielles à l'existence d'une fonction de méthodologie efficace, à savoir une offre de statisticiens ayant la formation voulue dans le pays, un bureau de la statistique doté d'une infrastructure qui fonctionne, des salaires qui, s'ils ne sont pas concurrentiels, s'approchent au moins de ceux offerts dans le secteur privé, et ainsi de suite). J'aborderai deux grands thèmes. L'un d'eux est la gestion de la tension entre l'indépendance et la pertinence : en général, les méthodologistes doivent collaborer étroitement avec le reste de l'organisme statistique pour que leurs travaux soient pertinents. En effet, ils doivent s'efforcer de servir les objectifs des clients externes, représentés au sein du bureau par les spécialistes des divers domaines. Cependant, pour être efficaces, les méthodologistes doivent





## Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'agencement de théorie et de pratique caractéristique des travaux de Waksberg.

Veillez consulter la section avis à la fin de la revue pour des informations sur le processus de nomination et de sélection du prix Waksberg 2012.

Ce numéro de *Techniques d'enquête* commence par le dixième article de la série du prix Waksberg. Le comité de rédaction remercie les membres du comité de sélection, composé de Leyla Mohadjer (Présidente), Daniel Kasprzyk, Elisabeth A. Martin et Wayne Fuller, d'avoir choisi Ivan P. Fellegi comme auteur de l'article du prix Waksberg de cette année.

### Article sollicité Waksberg 2010

Auteur : Ivan P. Fellegi

Ivan P. Fellegi est statisticien en chef émérite du Canada à Statistique Canada où il a été statisticien en chef de 1985 à 2008. C'est au cours de son mandat que Statistique Canada a reçu le titre de « meilleur organisme statistique au monde » de la revue *The Economist*. La contribution de M. Fellegi a été substantielle pour la méthodologie d'enquête, mais également pour la gestion performante d'un grand organisme au cours de sa longue carrière à Statistique Canada.

Monsieur Fellegi a beaucoup publié sur les méthodes statistiques, notamment sur les applications sociales et économiques des statistiques ainsi que sur la gestion fructueuse des organismes statistiques. Certains de ses articles ont marqué un tournant en méthodologie. Il a publié sur un bon nombre de sujets tels que le plan de sondage, le contrôle et l'imputation, le couplage de données et l'analyse des données d'enquête. De plus, il a participé activement à plusieurs comités. Il a été président de la Conférence des statisticiens européens de la Commission économique des Nations Unies pour l'Europe (1993-97), président du Comité sur les statistiques de l'Organisation de coopération et de développement économique (2004-2008), ancien président de l'Institut international de statistique, de l'Association internationale des statisticiens d'enquête et de la Société statistique du Canada et ancien président du Board of Governors, de l'Université Carleton (1995-97). Il a reçu de nombreux titres honorifiques et prix : Officier de l'Ordre du Canada, le Prix pour services insignés, de la fonction publique du Canada, l'Ordre du mérite de la République de Hongrie, le Prix pour carrière exceptionnelle de l'Initiative de recherche sur les politiques au Canada, la médaille de la Ville de Paris. Monsieur Fellegi est également membre de l'Académie des sciences de Hongrie. Il a reçu en outre la médaille d'or de la Société statistique du Canada et la médaille Robert Schuman de la Communauté européenne. Il est également récipiendaire de doctorats honorifiques de l'Université de Montréal, de l'Université du Québec (Institut national de la recherche scientifique), de l'Université Simon-Fraser, de l'Université McMaster, de l'Université d'Ottawa. De plus, il est membre honoraire de l'Institut international de la statistique et membre titulaire honoraire de la Royal Statistical Society.



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'«American National Standard for Information Sciences» – «Permanence of Paper for Printed Library Materials», ANSI Z39.48 - 1984.



**Techniques d'enquête**  
Une revue éditée par Statistique Canada  
Volume 36, numéro 2, décembre 2010

**Table des matières**

**Article Sollicite Waksberg**

Ivan P. Fellegi

L'organisation de la méthodologie statistique et de la recherche méthodologique  
dans les bureaux nationaux de la statistique.....

131

**Articles réguliers**

Carl-Erik Särndal et Sixten Lundström

Plan d'estimation : détermination de vecteurs auxiliaires en vue de réduire le biais de non-réponse.....

141

Jae Kwang Kim

Estimation par calage en utilisant l'inclinaison exponentielle dans les enquêtes par sondage.....

157

Stephen J. Haslett, Marissa C. Isidro et Geoffrey Jones

Comparaison de méthodes de régression sur données d'enquête dans le contexte de l'estimation  
de la pauvreté sur des petits domaines.....

169

Maria Rosaria Ferrante et Carlo Trivisano

Utilisation de modèles multivariés pour l'estimation sur petits domaines du nombre de recrues  
dans les entreprises.....

185

Julia D'Arrigo et Chris Skinner

Estimation de la variance par linéarisation pour les estimateurs par calage généralisé  
en présence de non-réponse.....

197

Abdellatif Demnati et J.N.K. Rao

Estimateurs de variance par linéarisation pour les paramètres de modèles à partir  
de données d'enquêtes complexes.....

211

Kirk M. Wolter, Phil Smith et Stephen J. Blumberg

Fondements statistiques des enquêtes par téléphone mobile.....

221

**Communications brèves**

Rudolf Witt, Diemuth E. Perns et Hermann Waibel

Collecte de données pour évaluer la pauvreté et la vulnérabilité dans les régions éloignées  
en Afrique subsaharienne.....

235

Mohammed G. Qayad, Pranesh Chowdhury, Shaohua Hu et Lina Balluz

Différences entre les répondants et durée de la période de collecte des données  
dans le Behavioral Risk Factor Surveillance System.....

241

Yves Tillé et David Haziza

Une propriété intéressante de l'entropie de certains plans d'échantillonnage.....

247

**Remerciements**

251

**Annonces**

253

**Autres revues**

255

# TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

*Techniques d'enquête* est répertoriée dans The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods. La revue est également citée par SCOPUS sur les bases de données Elsevier Bibliographic Databases.

## COMITÉ DE DIRECTION

Président

J. Kovar

Anciens présidents

D. Royce (2006-2009)

G.J. Brackstone (1986-2005)

R. Platak (1975-1986)

## COMITÉ DE RÉDACTION

Rédacteur en chef M.A. Hidiroglou, Statistique Canada

Rédacteur en

chef délégué H. Mantel, Statistique Canada

Rédacteurs associés

J.-F. Beaumont, Statistique Canada

J. van den Brakel, Statistics Netherlands

J.M. Brick, Westat Inc.

P. Cantwell, U.S. Bureau of the Census

R. Chambers, Centre for Statistical and Survey Methodology

J.L. Eltinge, U.S. Bureau of Labor Statistics

W.A. Fuller, Iowa State University

J. Gambino, Statistique Canada

B. Hülliger, University of Applied Sciences Northwestern Switzerland

D. Judkins, Westat Inc.

D. Kasprzyk, Mathematica Policy Research

P. Kott, National Agricultural Statistics Service

P. Lahiri, JPSM, University of Maryland

P. Lavallée, Statistique Canada

P. Lynn, University of Essex

D.J. Malec, U.S. Census Bureau

G. Nathan, Hebrew University

J. Opsomer, Colorado State University

D. Pfeffermann, Hebrew University

N.G.N. Prasad, University of Alberta

J.N.K. Rao, Carleton University

J. Reiter, Duke University

L.-P. Rivest, Université Laval

N. Schenker, National Center for Health Statistics

F.J. Schuren, National Opinion Research Center

P. do N. Silva, Escola Nacional de Ciências Estatísticas

P. Smith, Office for National Statistics

E. Stasny, Ohio State University

L. Stokes, Southern Methodist University

M. Thompson, University of Waterloo

V.J. Verma, Università degli Studi di Siena

K.M. Wolter, Iowa State University

C. Wu, University of Waterloo

W. Yung, Statistique Canada

A. Zaslavsky, Harvard University

Membres G. Beaudoin

S. Fortier (Gestionnaire de la production)

J. Gambino

M.A. Hidiroglou

H. Mantel

Ancien rédacteur en chef

J. Kovar (2006-2009)

M.P. Singh (1975-2005)

## POLITIQUE DE RÉDACTION

*Techniques d'enquête* publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

## Présentation de textes pour la revue

*Techniques d'enquête* est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférentiellement en Word au rédacteur en chef. (mailto:statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca).

## Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada : États-Unis 12 \$ CA (6 \$ x 2 exemplaires), autres pays, 20 \$ CA (10 \$ x 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.



# Techniques d'enquête

18446

Décembre 2010  
N° 12-001-XPB au catalogue  
Périodicité : semestrielle  
ISSN 0714-0045  
Ottawa

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2010  
Publication autorisée par le ministre responsable de Statistique Canada

Décembre 2010 • Volume 36 • Numéro 2

Une revue  
éditée  
par Statistique Canada



Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à [www.statcan.gc.ca](http://www.statcan.gc.ca). Vous pouvez également communiquer avec nous par courriel à [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca) ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements  
Service national d'appareils de télécommunications pour les malentendants  
1-800-263-1136  
1-800-363-7629  
1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements  
Télécopieur  
1-613-951-8116  
1-613-951-0581

Programme des services de dépôt

Service de renseignements  
Télécopieur  
1-800-635-7943  
1-800-565-7757

Comment accéder à ce produit ou le commander

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à [www.statcan.gc.ca](http://www.statcan.gc.ca) et de parcourir par « Ressource clé » > « Publications ».

Ce produit est aussi disponible en version imprimée standard au prix de 30 \$CAN l'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 \$CAN	12 \$CAN
Autres pays	10 \$CAN	20 \$CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca)
- Poste
- En personne auprès des agents et librairies autorisés.

Finances  
Immeuble R.-H.-Coats, 6<sup>e</sup> étage  
150, promenade Tunney's Pasture  
Ottawa (Ontario) K1A 0T6

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site [www.statcan.gc.ca](http://www.statcan.gc.ca) sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

---

# Techniques d'enquête

---

N° 12-001-XPB au catalogue

Une revue  
éditée  
par Statistique Canada

Décembre 2010

•

Volume 36

•

Numéro 2



Statistique  
Canada

Statistics  
Canada

Canada











